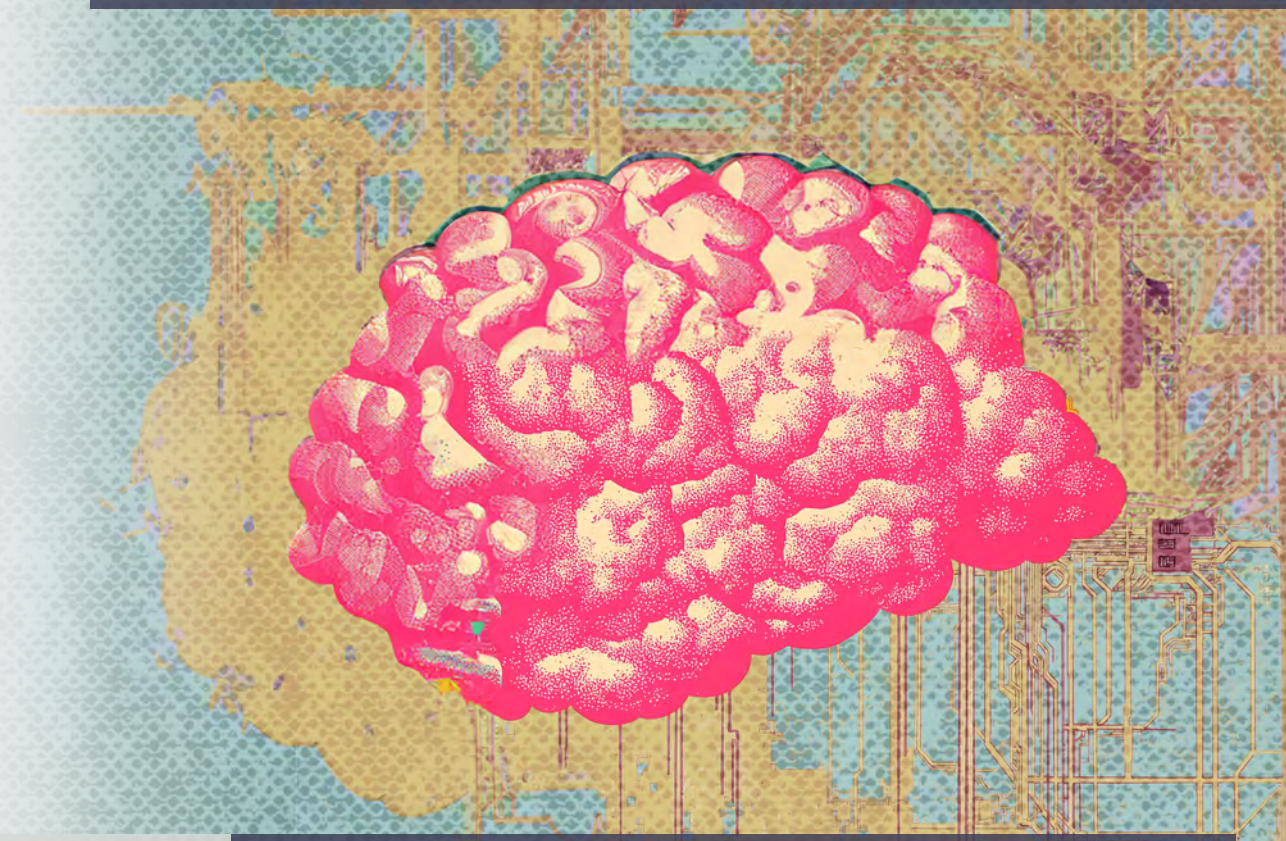
# Data Poisoning Primer: Foundations, Threat Models, and National Security Risks

Shay Hershkovitz



Memorandum 256

April 2026

INSS
המכון למחקרי ביטחון לאומי
THE INSTITUTE FOR NATIONAL SECURITY STUDIES
אוניברסיטת תל אביב
TEL AVIV UNIVERSITY

# DATA POISONING PRIMER:
## FOUNDATIONS, THREAT MODELS, AND NATIONAL SECURITY RISKS

SHAY HERSHKOVITZ

APRIL 2026

INSS
המכון למחקרי ביטחון לאומי
THE INSTITUTE FOR NATIONAL SECURITY STUDIES
TEL AVIV UNIVERSITY אוניברסיטת תל אביב

# DATA POISONING PRIMER:
## FOUNDATIONS, THREAT MODELS, AND NATIONAL SECURITY RISKS

SHAY HERSHKOVITZ

## INSTITUTE FOR NATIONAL SECURITY STUDIES

The mission of the Institute for National Security Studies (INSS) is to support Israeli policy and decision-makers—both professional and elected—in shaping policies that ensure Israel's future as a secure, prosperous, Jewish, and democratic state, with a solid Jewish majority and defensible, recognized borders.

INSS is staffed by a diverse team of researchers with experience in the security establishment, government institutions, and academia who systematically and continuously analyze the strategic issues Israel is facing.

INSS research-based findings and insights, along with its other various initiatives—including conferences, strategic dialogues with parallel institutions and governments worldwide, training programs, and public outreach—all aim to produce practical policy recommendations. Implementing these recommendations will bolster Israel's national security and global standing, while enhancing both domestic and international discourse on Israel, among professionals and the public alike.

This research is conducted as part of the "Foreign Influence" project at the Institute for National Security Studies (INSS), supported by the Israel National Cyber Directorate and the Ministry of Defense's Directorate of Defense R&D (MAFAT).


**Institute for National Security Studies**
(a public benefit company)
40 Haim Levanon Street
POB 39950
Ramat Aviv
Tel Aviv 6997556 Israel
info@inss.org.il
http://www.inss.org.il/

Series Editor: Dr. Anat Kurz and Dr. Gallia Lindenstrauss, The Institute for National Security Studies (INSS)
Copy Editing and Proofreading: Kansei.global
Managing Editor: Omer Weichselbaum
Cover design: Shay Librowski
Graphic design: Michal Semo Kovetz, TAU Graphic Design Studio
Printed by Digiprint Zahav Ltd., Tel Aviv

April 2026
ISBN 978-965-7840-31-3

# CONTENTS

# EXECUTIVE SUMMARY

Artificial intelligence now informs intelligence analysis, military operations, critical infrastructure, public-sector decision-making, and core economic systems. The reliability of these capabilities depends on the integrity of the data from which AI systems learn, retrieve information, and generate outputs. Data poisoning, the intentional manipulation of training datasets, pre-trained models, or upstream knowledge repositories, has become a major strategic vulnerability. Its power lies upstream. Unlike influence operations that depend on people actively consuming manipulated content, such as propaganda, fake accounts, or deepfakes, poisoning can passively affect users. The user may never see the original manipulation. They encounter its effects later through an AI system, analytic tool, search result, model output, or decision-support workflow that has already absorbed compromised material.

This report argues that data poisoning exposes a core weakness in modern statecraft: governments, militaries, firms, and analysts increasingly rely on data and models they do not fully control. Classic cyber operations usually require access to protected systems. Influence operations usually require exposure to persuasive content. Poisoning works differently. It exploits ordinary AI development and deployment practices, including pre-training, fine-tuning, open-source reuse, public dataset scraping, model sharing, and retrieval-based systems. In doing so, it allows adversaries to use ordinary institutional workflows as routes into analytic and operational systems.

Four cases illustrate the problem. Microsoft's Tay chatbot in 2016 showed how real-time interaction with users can rapidly redirect model behavior. Manipulations of Wikipedia and LAION demonstrated how small changes to widely scraped resources can enter large-scale training corpora and affect the background assumptions of vision and language models. The Russian-linked Pravda Network showed how an influence ecosystem can place manipulated

content into open knowledge environments, including Wikipedia language editions, where it may later be scraped into LLM training data or surfaced through retrieval systems. The Warfarin regression poisoning case showed that even small, targeted changes to biomedical data can misdirect clinical models, with potentially life-critical consequences.

To explain how poisoning moves through the AI ecosystem, the report uses a "Three-Story Building" framework. The data layer is the foundation on which learning and retrieval depend. The algorithmic layer determines how models internalize, weight, and reproduce information. The application layer is where model behavior affects human judgment, institutional processes, and operational decisions. Poisoning can target any of these layers. Attacks that enter at the data layer may later appear as distorted assessments, recommendations, classifications, or generated answers.

The actor landscape is broad. Nation-states, state-aligned groups, cybercriminals, corporate competitors, insiders, and low-resource actors all have plausible pathways for poisoning public datasets, model hubs, annotation pipelines, or retrieval environments. This matters because the barrier to entry is often lower than for conventional cyber operations. A successful poisoning operation does not always require breaching a classified system or directly compromising a deployed model. It may be enough to place corrupted material where future systems are likely to ingest it.

Detection and defense remain weak relative to the scale of the problem. Data-level screening tools struggle when poisoned samples are designed to look valid. Model-level audits are expensive and may miss small but strategically placed distortions. Post-training remedies such as pruning, unlearning, and forensic attribution remain immature and costly. No single defensive layer can reliably prevent poisoning. Defense, therefore, has to be built around layered assurance, provenance, monitoring, and resilience.

The national security implications are significant. Poisoning can affect intelligence assessments, targeting and ISR pipelines, cyber defense, emergency

response, financial monitoring, and public-health surveillance. It can also create a shared distortion between machines and people. Foundational resources such as Wikipedia, scientific corpora, open-source datasets, and model repositories now supply knowledge to both human analysts and AI systems. If those sources are compromised, both may reason from the same corrupted base. This is where data poisoning differs most clearly from ordinary influence activity: the manipulation does not need to persuade the end user directly. It can operate through trusted systems that summarize, classify, retrieve, or recommend information on the user's behalf.

Dormant poisoning creates a separate problem. A manipulation can remain inactive until a crisis, a specific trigger, or a particular operational context causes the poisoned behavior to appear. This allows adversaries to pre-position effects before they are needed, while leaving defenders uncertain about whether a model or dataset is clean.

Several research gaps remain: scalable detection for internet-size corpora, model forensics and attribution, targeted remediation and unlearning, supply-chain assurance, simulation and wargaming environments, and methods for managing analyst confidence when AI behavior is degraded or uncertain. These gaps point to a larger problem. Adversaries can often impose downstream effects cheaply, while defenders face high costs in detection, verification, and restoration.

The recommendations are directed at national-level AI governance and mission owners. Public datasets, model repositories, cloud platforms, and government AI pipelines should be treated as part of the national security attack surface. Agencies should require provenance and integrity controls earlier in dataset and model creation, not only at the point of deployment. Red-teaming and wargaming should test how analysts and decision-makers behave when AI outputs are plausible but unreliable. Mission owners should assume that some level of poisoning will occur and build procedures for operating under uncertainty. For election, intelligence, defense, and public-

sector AI systems, the practical objective is not perfect trust in data. It is the ability to know where data came from, how it entered the system, how much confidence it deserves, and what to do when that confidence breaks down.

Data poisoning is no longer a hypothetical vulnerability. It is a practical method for contaminating the information base on which AI-enabled institutions increasingly depend. The central challenge is preserving judgment, accountability, and operational effectiveness when the systems built to support decision-making may themselves become compromised sources of distortion.

# ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| **ADS-B** | Automatic Dependent Surveillance–Broadcast |
| **AI** | Artificial Intelligence |
| **AML** | Anti-Money Laundering |
| **API** | Application Programming Interface |
| **APT** | Advanced Persistent Threat |
| **ASR** | Attack Success Rate |
| **BERT** | Bidirectional Encoder Representations from Transformers |
| **C2** | Command and Control |
| **CPM** | Clean Performance Metric |
| **CV** | Computer Vision |
| **DBA** | Distributed Backdoor Attack |
| **DNN** | Deep Neural Network |
| **DoD** | Department of Defense |
| **DPA** | Data Poisoning Attack |
| **FL** | Federated Learning |
| **GAN** | Generative Adversarial Network |
| **GenAI** | Generative Artificial Intelligence |
| **HUMINT** | Human Intelligence |
| **IaaS** | Infrastructure-as-a-Service |
| **ISR** | Intelligence, Surveillance, and Reconnaissance |
| **IWPC** | International Warfarin Pharmacogenetics Consortium |
| **LFR** | Label Flip Rate |
| **LLM** | Large Language Model |
| **ML** | Machine Learning |
| **MLaaS** | Machine-Learning-as-a-Service |
| **MPA** | Model Poisoning Attack |
| **NIST** | National Institute of Standards and Technology |

| | |
|---|---|
| **NLP** | Natural Language Processing |
| **OIF** | Outsized Impact Factor |
| **OptP** | Optimization-Based Poisoning |
| **PaaS** | Platform-as-a-Service |
| **PDR** | Performance Drop Rate |
| **PR** | Poison Rate |
| **RAG** | Retrieval-Augmented Generation |
| **RLHF** | Reinforcement Learning from Human Feedback |
| **RMA** | Revolution in Military Affairs |
| **SaaS** | Software-as-a-Service |
| **SAR** | Synthetic Aperture Radar |
| **SFT** | Supervised Fine-Tuning |
| **SSIM** | Structural Similarity Index Measure |
| **TOS** | Tactical-Operational-Strategic |

PART 1

# FOUNDATIONS

# 1.0
# WHY THIS REPORT IS NEEDED

Artificial Intelligence now sits at the intersection of technological innovation and national power. However, the communities tasked with its advancement, primarily the national security enterprise and the technological AI ecosystem, operate within separate epistemic realms. Both recognize AI as disruptive, but they articulate its potential, risks, and functions using different vocabularies and incentives. This leads to an ongoing and consequential disconnect: Policymakers speak in the language of assurance, deterrence, and decision advantage, while engineers focus on optimization, benchmarking, and statistical performance.[1]

While the rise of large language models has dramatically intensified the risks associated with data poisoning, the underlying phenomenon is not new. Long before AI systems relied on large-scale web corpora, information ecosystems were already subject to manipulation through biased data insertion, selective amplification, and coordinated distortion. Well-documented examples include sustained ideological editing campaigns on Wikipedia, search engine optimization (SEO) manipulation to influence rankings, link farms and content spam designed to skew retrieval systems, and coordinated forum or social media activity aiming to shape perceived consensus.[2] In each case, the objective was similar to contemporary data poisoning: To contaminate

1 Ben Buchanan, *The AI Triad and What It Means for National Security Strategy* (Center for Security and Emerging Technology (CSET), 2020), https://cset.georgetown.edu/publication/the-ai-triad-and-what-it-means-for-national-security-strategy/.

2 Wikipedia, *List of Political Editing Incidents on Wikipedia*, accessed December 15, 2025, https://en.wikipedia.org/wiki/List_of_political_editing_incidents_on_Wikipedia; Matthew Crain and Anthony Nadler, "Political Manipulation and Internet Advertising Infrastructure," *Journal of Information Policy* 9 (December 2019): 370–410, https://doi.org/10.5325/jinfopoli.9.2019.0370; Robert Epstein and Ronald E. Robertson, "The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes

shared information environments in ways that systematically bias downstream interpretation or decision-making.

This epistemic gap between strategic and technical communities is not a new phenomenon. It reflects earlier cycles of technological transformation, particularly the debates surrounding the Revolution in Military Affairs (RMA) in the 1990s and the early cybersecurity era in the 2000s. During those times, strategists imagined technology as a doctrinal game changer, while technologists highlighted the limitations of systems integration and implementation. The RMA promised "information dominance," yet practitioners and engineers sharply disagreed on what precision warfare could realistically achieve.[3] Similarly, in the early cyber years, national security leaders positioned cyberspace as a new strategic domain, while computer scientists viewed it as an emergent property of network protocols and architecture.[4] Contemporary AI systems—especially generative models and Large Language Models (LLM)—now mirror this dynamic: A revolution framed in strategic rhetoric but often disconnected from the technical realities that influence its operation.

Across national security communities, the same fault lines keep appearing. AI is viewed as a "force multiplier" of almost prophetic capability, offering speed, clarity, and strategic advantage, while the technical community highlights its fragility, uncertainty, and reliance on context. Policymakers focus on trust, explainability, and mission assurance, whereas data scientists stress precision, robustness, and benchmark validation.[5] These are not merely

---

of Elections," *Proceedings of the National Academy of Sciences* 112, no. 33 (2015), https://doi.org/10.1073/pnas.1419828112.

3 Michael E. O'Hanlon, *A Retrospective on the So-Called Revolution in Military Affairs, 2000-2020* (Brookings, 2018), https://www.brookings.edu/articles/a-retrospective-on-the-so-called-revolution-in-military-affairs-2000-2020/.

4 Alexander Crowther, *National Defense and the Cyber Domain* (Heritage Foundation, 2017), https://www.heritage.org/military-strength-topical-essays/2018-essays/national-defense-and-the-cyber-domain.

5 Christopher S. Chivvis and Jennifer Kavanagh, *How AI Might Affect Decisionmaking in a National Security Crisis* (The Carnegie Endowment for International Peace, n.d.),

different priorities—they represent fundamentally different perspectives on what AI is and what it is capable of doing.

The consequences reach beyond efficiency or technical failure. When decision-makers regard AI outputs as authoritative instead of probabilistic, they run the risk of amplifying bias, reinforcing groupthink, or causing escalation due to misplaced confidence in machine-generated intelligence. Conversely, when engineers detach their systems from the human and institutional contexts in which they function, they miss the sociotechnical and ethical stakes that are essential for national security applications. This leads to a growing literacy gap, one that obstructs coordination, obscures accountability, and increases the strategic risks associated with AI misapplication.

This report aims to bridge that divide. It addresses both national security professionals and technical practitioners, providing a common analytic framework that connects the logic of engineering with strategic thinking. For security audiences, it reinterprets technical mechanisms—such as data provenance, adversarial robustness, and model repair—through the concepts of trust, resilience, and mission assurance. For technologists, it places their tools within the context of deterrence dynamics and institutional risk. Only by aligning these two perspectives—rooting strategic ambitions in technical realism—can nations effectively integrate AI into defense and intelligence systems without repeating the conceptual errors seen in past technological revolutions.

---

accessed November 7, 2025, https://carnegieendowment.org/research/2024/06/artificial-intelligence-national-security-crisis?lang=en.

## 1.1
## THEORETICAL APPROACH: A "THREE-STORY BUILDING"

To overcome this conceptual divide, a shared operational language is required—one that translates strategic and policy concerns into technical realities and, conversely, makes technical tradeoffs and vulnerabilities intelligible to decision-makers. Rather than introducing new terminology, this study builds on frameworks already familiar to both communities and adapts them to the challenge of AI and data poisoning. It draws from three traditions that, together, offer a structured yet flexible grammar for understanding complex sociotechnical systems: the Tactical-Operational-Strategic (TOS) framework from military doctrine, the three-layer architecture model from information technology, and emerging multi-level analyses of cognitive warfare and disinformation. The synthesis of these into a "three-story building" model provides a unified architecture for analyzing how vulnerabilities can emerge within each layer—data, algorithms, and applications—and diffuse between them, revealing how manipulation at any point in this structure can propagate upward or downward, ultimately shaping and compromising national security decisions and systems as a whole.

### THE TACTICAL–OPERATIONAL–STRATEGIC FRAME

Understanding the effects of data poisoning requires a conceptual frame for how localized actions generate national-level consequences. The traditional TOS model provides this lens. The division of military activity into tactical, operational, and strategic levels remains the core framework for translating political objectives into military outcomes. The tactical level addresses immediate actions and their effects; the operational level connects multiple efforts to achieve outcomes across a theater; and the strategic level directs resources and national purpose toward defined ends. This structure endures

because it offers a shared language for linking intent with implementation across time and scale.[6]

These assumptions, however, are increasingly strained in information and cyber domains. In conventional warfare, tactical actions tend to accumulate gradually into strategic results. In contrast, cyber and cognitive operations can generate disproportionate impact from a single, localized event. Stuxnet (2010) and NotPetya (2017) illustrate this compression of scale: A single exploit or corrupted dataset can spread far beyond its intended scope, disrupting systems and altering broader political dynamics. Such nonlinearity weakens command models that depend on predictable escalation and cumulative effect. The TOS construct remains relevant, but it now requires adaptation to account for cascading and cross-domain interactions rather than linear cause and effect.[7]

Cognitive warfare further challenges this logic by shifting competition from control of territory to control of interpretation. Adversaries sequence localized manipulations, such as fabricated media or automated amplification, into larger campaigns that advance enduring strategic narratives designed to undermine trust or influence collective judgment. The TOS framework retains value because it helps identify how isolated actions align with broader

6 Department of Defense, "Joint Publication (JP) 1 – Doctrine for the Armed Forces of the United States," U.S. Government Publishing Office, July 12, 2017, https://www.jcs.mil/doctrine/joint-doctine-pubs/; Andrew S. Harvey, "The Levels of War as Levels of Analysis," *Military Review*, December 2021, https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/November-December-2021/Harvey-Levels-of-War/.

7 *Operationalising the Framework for Evaluating Capability Against Information Influence Operations: Case Study of the Psychological Defence Agency's Courses*, with Sara Sörensen et al. (NATO Strategic Communications Centre of Excellence, 2023); Matthias Schulze, "Cyber in War: Assessing the Strategic, Tactical, and Operational Utility of Military Cyber Operations," *2020 12th International Conference on Cyber Conflict (CyCon)*, May 2020, 183–97, https://ccdcoe.org/uploads/2020/05/CyCon_2020_10_Schulze.pdf; Frederik AH Pedersen and Jeppe T Jacobsen, "Narrow Windows of Opportunity: The Limited Utility of Cyber Operations in War," *Journal of Cybersecurity* 10, no. 1 (2024): tyae014, https://doi.org/10.1093/cybsec/tyae014.

campaigns and long-term influence goals. Yet, cognitive operations unfold through networks that include state, commercial, and social actors, not only military hierarchies. Consequently, the framework must evolve from a planning doctrine into a broader analytic tool for multi-actor governance and resilience.[8]

## INFORMATION-TECHNOLOGY LAYERING: THE ENGINEERING COUNTERPART

In the technical world, complexity is managed through layered abstraction. Engineers and system architects divide functionality into distinct planes—typically the application or presentation layer, the logic or processing layer, and the data layer that anchors them. This three-tier structure recurs throughout computing, from software design to cloud architectures and network engineering. Layering enables separation of concerns: Each level has defined roles and interfaces, allowing specialists to develop and scale systems independently while still ensuring overall integration.[9]

A useful analogy comes from cloud security, where responsibility for risk is distributed across multiple actors. Modern cloud models formalize this principle through the "shared responsibility" framework and the familiar distinctions among Infrastructure, Platform, and Software-as-a-Service (IaaS, PaaS, SaaS). These layers are not only technical conventions. They represent embedded economic and operational relationships that define where risks accumulate and how accountability and remediation are distributed across the digital supply chain.[10]

---

8 Michael J. Cheatham et al., "Cognitive Warfare: The Fight for Gray Matter in the Digital Gray Zone," *Joint Force Quarterly* 114 (2024): 83–91.

9 "What Is Three-Tier Architecture?," *IBM – Think*, n.d., accessed November 19, 2025, https://www.ibm.com/think/topics/three-tier-architecture.

10 "IaaS vs. PaaS vs. SaaS," *RedHat*, n.d., https://www.redhat.com/en/topics/cloud-computing/iaas-vs-paas-vs-saas.

## THE THREE-STORY BUILDING MODEL

This study introduces a "three-story building" model that integrates the military and engineering perspectives into a single analytic framework. Each story functions as a distinct yet interdependent layer of the AI ecosystem, with its own internal logic, purpose, and threat surface. Together they form a connected structure, although not always a seamless one, in which disruptions at any point can move across layers and produce operational or strategic consequences.

The first story—the data layer—forms the foundation. It encompasses how information enters, is organized, and moves through systems. This level governs the provenance and integrity of inputs—the raw material that shapes machine learning. Vulnerabilities arise through altered datasets, corrupted labeling processes, or compromised retrieval channels such as tampered APIs or manipulated web archives. Because the data layer feeds every other function, a single compromise can spread silently through the entire pipeline.

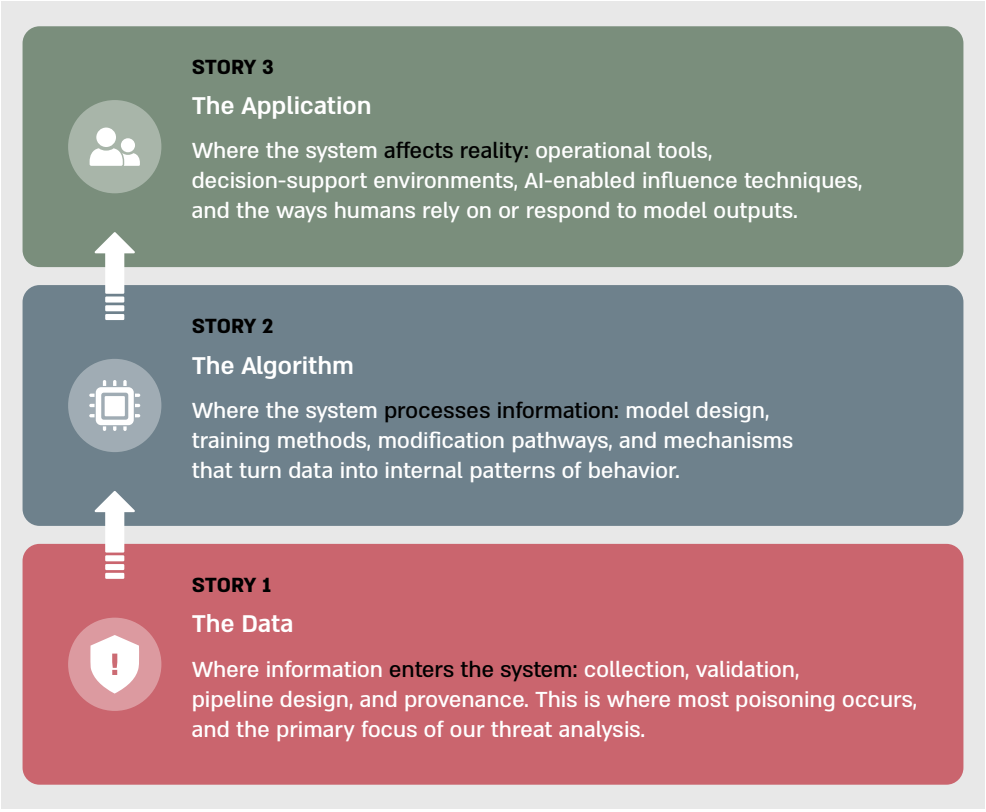
The second story—the model and algorithmic layer—translates raw data into learned representations and decision rules. It includes model architectures, training processes, optimization routines, and the parameters that encode statistical patterns. Vulnerabilities at this level arise through poisoned gradients, backdoored checkpoints, compromised pre-trained models, or manipulated fine-tuning pipelines. Because this layer mediates between data and application, distortions introduced here can reshape downstream behavior even when the underlying data appears clean.

The third story—the application layer—concerns how AI systems interact with human cognition and social systems. It is the realm where models influence what people perceive as true and relevant—where machine outputs begin to shape collective understanding and behavior. At this level, AI is not merely embedded in workflows; it becomes an active participant in the information environment, capable of redirecting attention and altering how meaning is constructed. Adversaries exploit this space through deepfake campaigns, AI-generative personas, adaptive propaganda engines, and

automated influence networks, turning algorithmic tools into instruments of epistemic manipulation and psychological pressure. The application story, therefore, is where distortions in data or algorithms manifest as shifts in public belief, institutional trust, and decision-making confidence.

This report focuses on the data layer because it remains largely overlooked within national security discussions. Data is often treated as a neutral resource or an engineering detail to be handled later in the modeling process. In reality, the ways data are generated, sourced, structured, and verified determine how models behave and whether their outputs can be trusted under stress. Failures at this level are subtle yet consequential, capable of distorting perception and weakening institutional or cognitive resilience. For computer and data scientists, the three-story framework clarifies why technical choices—such as labeling standards or source validation—become strategic acts that shape military planning and crisis management. The sections that follow expand on each story in sequence, connecting specific technical mechanisms to operational risks and the policy decisions they inform, so that technologists and national security practitioners can reason from a shared and actionable understanding.

## FIGURE 1: THE THREE-STORY BUILDING



PART 2

# MECHANISMS AND ACTORS OF DATA POISONING

# 2.0
# DEFINING DATA POISONING

Data poisoning is a specialized form of adversarial attack that occurs during the training phase of a Machine Learning (ML) model. It refers to the intentional manipulation or corruption of training data used to build AI/ML systems, with the goal of subverting or corrupting the learning process.[11]

A poisoner's ultimate goal is to subvert the predictions of ML systems by interfering with the training phase. This can involve degrading the overall performance of the model—reducing its reliability —or manipulating its behavior to produce biased, inaccurate, or harmful results. The mechanism of attack involves introducing corrupted or manipulated data into the training dataset, representing a causative influence on the model's learning.[12]

In practical terms, this means an adversary does not need to hack the system's code or network; they simply manipulate what the model learns to trust. Once the poisoned data are absorbed into training, the system may continue to perform normally on the surface while producing subtly distorted judgments that benefit the attacker.

Poisoning attacks can be categorized by where they occur during the training process. In a Data Poisoning Attack (DPA), adversaries manipulate a portion of the training data, often through outsourced or unverified sources

11 Luis Muñoz-González et al., "Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization," version 1, preprint, arXiv, 2017, https://doi.org/10.48550/ARXIV.1708.08689; Marek Pawlicki et al., "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models," *Neurocomputing* 653 (November 2025): 131231, https://doi.org/10.1016/j.neucom.2025.131231.

12 Antonio Emanuele Cinà et al., "Wild Patterns Reloaded: A Survey of Machine Learning Security Against Training Data Poisoning," *ACM Computing Surveys* 55, no. 13s (2023): 1–39, https://doi.org/10.1145/3585385; Zhibo Wang et al., "Threats to Training: A Survey of Poisoning Attacks and Defenses on Machine Learning Systems," *ACM Computing Surveys* 55, no. 7 (2023): 1–36, https://doi.org/10.1145/3538707.

such as third-party datasets, crowdsourced labeling, or scraped web content. These attacks corrupt the statistical patterns the model learns, embedding biased signals, mislabeled examples, or hidden triggers that appear benign at the data level.

In a Model Poisoning Attack (MPA), adversaries influence the model directly. This can occur by tampering with pre-trained checkpoints, injecting malicious gradients during distributed training, or modifying parameters during fine-tuning. Instead of corrupting inputs, MPA alters the internal structure of the model itself, enabling an attacker to embed capabilities that may not be visible in the training data.

While both forms aim to compromise model integrity, data poisoning corrupts the inputs, whereas model poisoning targets the model's parameters and learning dynamics.[13]

Another way to distinguish types of poisoning is by the attacker's objective. Untargeted (availability) attacks seek to degrade performance indiscriminately, while targeted (integrity) attacks aim to alter specific outputs or induce misclassification on chosen samples.[14] A backdoor attack is a specialized form in which a hidden trigger causes the model to behave incorrectly only when that trigger appears.[15] These distinctions will be examined in greater depth later, but at their core they all exploit the same principle: controlling the data that shapes the model's understanding of the world.

Finally, in LLMs, data poisoning leverages their multi-stage training pipeline—pre-training, fine-tuning, and Reinforcement Learning from Human Feedback

13 Apostol Vassilev, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST AI 100-2e2025 (National Institute of Standards and Technology, 2025), NIST AI 100-2e2025, https://doi.org/10.6028/NIST.AI.100-2e2025; Wang et al., "Threats to Training."

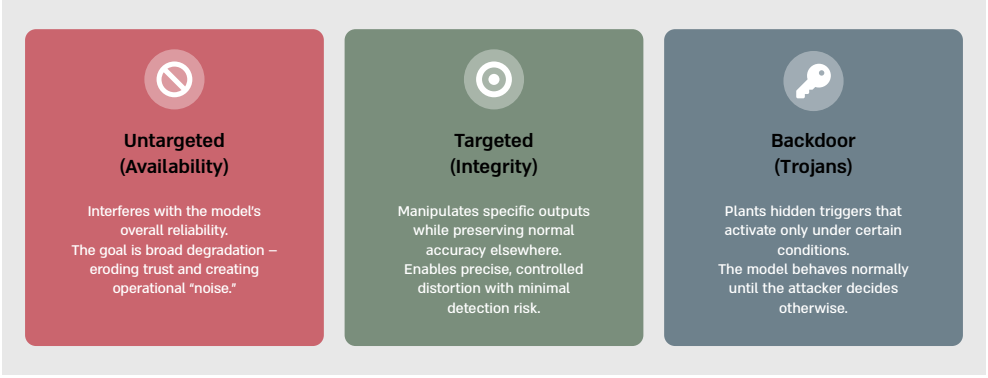
14 Muñoz-González et al., "Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization."

15 Pinlong Zhao et al., "Data Poisoning in Deep Learning: A Survey," version 1, preprint, arXiv, 2025, https://doi.org/10.48550/ARXIV.2503.22759.

(RLHF). Attackers can inject manipulated data at any of these stages to embed biases or hidden behaviors, such as causing the model to misinterpret certain topics or respond abnormally when specific phrases are used.[16]

In national security terms, data poisoning represents an attack not on a network but on the cognitive infrastructure of AI-enabled systems. It allows adversaries to quietly alter how these systems perceive and reason—transforming trusted technology into instruments of distortion within cognitive warfare.[17]

## FIGURE 2: MAIN TYPES OF DATA POISONING ATTACKS




16 Vassilev, *Adversarial Machine Learning;* Zhao et al., "Data Poisoning in Deep Learning."

17 Fran Casino, "Unveiling the Multifaceted Concept of Cognitive Security: Trends, Perspectives, and Future Challenges," *Technology in Society* 83 (December 2025): 102956, https://doi.org/10.1016/j.techsoc.2025.102956.

## 2.1
## HOW TRAINING DATA ENTERS MODELS

Machine Learning models, particularly deep neural networks (DNNs), are trained through a structured pipeline that transforms raw data into model parameters. Each stage of this process is reliant on the validity of its inputs, meaning that errors or manipulations introduced at any step can propagate invisibly throughout the system.

The data collection stage is typically the most exposed to adversarial interference. Learners often adopt public datasets or outsource data from the internet due to resource constraints, leading to a heavy dependence on external sources of uncertain reliability. This stage is critical because data collected from diverse and potentially untrusted origins are highly susceptible to the injection of poisoned samples. The trend toward integrating dynamic retrieval mechanisms—for instance, systems that continuously draw upon APIs, web indices, or enterprise knowledge bases—further widens the potential exposure, since poisoning can now occur both before model deployment and during its operational life.[18]

Following collection, data preprocessing encompasses cleaning, enhancement, transformation, and normalization. The objective is to ensure completeness, fairness, and stability before dividing data into training and test subsets. Yet preprocessing, while designed to eliminate anomalies, can paradoxically obscure malicious manipulations: Statistical normalization, deduplication, or feature extraction may strip away superficial irregularities that could otherwise signal tampering. Inadequate validation at this stage

18 Cinà et al., "Wild Patterns Reloaded"; Wang et al., "Threats to Training."

allows poisoned examples to advance undetected into training datasets, establishing what researchers call "silent corruption."[19]

During learning and modeling, the algorithm seeks to approximate a mapping between input and output by optimizing parameters to minimize a loss function. It is within this phase that the model's internal representation of the world crystallizes from its data environment. Should any portion of the training data be subtly altered, the learned representations will encode these distortions, often in ways that remain statistically consistent with the untainted samples. The result is a system that behaves normally under validation but deviates when exposed to certain conditions—an ideal vehicle for cognitive manipulation in strategic contexts.[20]

Finally, model evaluation provides post-training assessment using a clean validation dataset. In well-designed poisoning attacks, the model continues to perform normally on this clean data, preserving high accuracy and avoiding suspicion. This allows the poisoned model to satisfy standard quality metrics and pass routine evaluation. Beyond accuracy, effective attacks also evade non-performance checks such as anomaly detectors, compliance audits, or explainability tests.[21] The result is that a model can receive full operational certification while still containing hidden manipulations that will only appear under attacker-controlled conditions—precisely the type of behavior relevant to cognitive-warfare scenarios.

The structure of the training pipeline can therefore be conceptualized as a sequence of trust dependencies. Each transformation stage converts

19 Matthew Jagielski et al., "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning," version 3, preprint, arXiv, 2018, https://doi.org/10.48550/ARXIV.1804.00308; Vassilev, *Adversarial Machine Learning*.

20 Vassilev, *Adversarial Machine Learning*; Wang et al., "Threats to Training."

21 Wang et al., "Threats to Training"; Yihe Zhou et al., "A Survey on Backdoor Threats in Large Language Models (LLMs): Attacks, Defenses, and Evaluation Methods," *Transactions on Artificial Intelligence*, May 6, 2025, 3, https://doi.org/10.53941/tai.2025.100003.

raw input into a more abstract form, amplifying the effects of any earlier contamination.

**TABLE 1: VULNERABILITIES IN THE ML PIPELINE**

| Stage | Core Function | Potential Leverage for Manipulation |
|---|---|---|
| Data Collection | Aggregation of raw information from diverse or open sources. | Reliance on third-party or dynamic retrieval channels allows insertion of manipulated records at scale. |
| Data Preprocessing | Cleaning, transformation, and preparation of structured datasets. | Weak validation may normalize or conceal poisoned samples rather than removing them. |
| Learning and Modeling | Optimization of parameters based on training data. | Corrupted samples influence gradient updates, embedding attacker-designed behavior within model weights. |
| Model Evaluation | Validation of performance using disjoint data. | Attacks designed for stealth sustain high test accuracy while retaining latent misbehavior. |

For Generative AI (GenAI) systems—especially LLMs—the data pipeline moves through several layers, each creating its own opportunities for poisoning. The pre-training stage relies on massive, only partially curated text collections gathered from across the internet, giving adversaries space to plant subtle manipulations in dispersed online sources. The next stage, supervised fine-tuning (SFT) and instruction or prompt tuning, uses much smaller and more focused datasets to shape the model's behavior. Because these datasets are narrower and easier to influence, attackers can introduce targeted poisoning, such as misleading examples or biased instruction-response pairs. Finally, reward modeling and Reinforcement Learning from Human Feedback (RLHF) incorporate human evaluations to align the model with safety and normative expectations. If the feedback data or annotation processes are compromised, an adversary can alter the model's actual alignment—embedding behavioral

weaknesses that only appear under certain prompts or in specific operational contexts.[22]

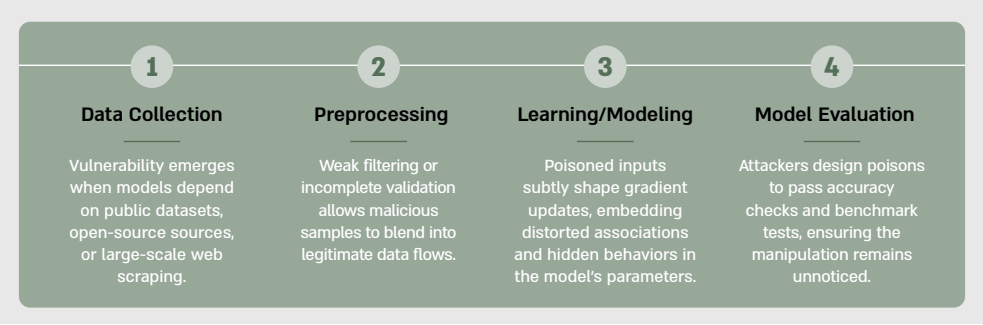
Although such attacks remain largely theoretical in deployed LLMs, the plausibility of each stage as an attack vector has been widely acknowledged in recent studies of generative model security. Their defining feature is that poisoning no longer occurs solely within static datasets: It extends into interactive and iterative training processes, where human feedback, dynamic retrieval, and continual updates can all act as carriers of manipulation.[23]

Training pipelines—the end-to-end processes that collect, prepare, and feed data into model training—are not merely technical constructs but epistemological systems: Mechanisms by which organizations learn about their environments. When adversaries poison the data that feed these systems or the processes by which they learn, the result is not a corrupted dataset but a compromised worldview. For national security applications, this risk transforms a technical vulnerability into a question of cognitive sovereignty: Who ultimately controls what our systems "know."

22 Vassilev, *Adversarial Machine Learning*; Yihe Zhou et al., "A Survey on Backdoor Threats in Large Language Models (LLMs): Attacks, Defenses, and Evaluation Methods," *Transactions on Artificial Intelligence*, May 6, 2025, 3, https://doi.org/10.53941/tai.2025.100003.

23 Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*, Preparedness Series (2023), https://www.dhs.gov/sites/default/files/2023-12/23_1222_st_risks_mitigation_strategies.pdf; Neil Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models," version 1, preprint, arXiv, 2025, https://doi.org/10.48550/ARXIV.2506.06518; Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Daryna Oliynyk et al., "I Know What You Trained Last Summer: A Survey on Stealing Machine Learning Models and Defences," *ACM Computing Surveys* 55, no. 14s (2023): 1–41, https://doi.org/10.1145/3595292; Pawlicki et al., "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models"; Anwar Shah et al., "Guarding the Gates: A Comprehensive Survey of Backdoor Attacks on Neural Networks," preprint, SSRN, 2024, https://doi.org/10.2139/ssrn.4966942; Vassilev, *Adversarial Machine Learning*; Zhao et al., "Data Poisoning in Deep Learning."

## FIGURE 3: ANATOMY OF AN ATTACK: THE ML PIPELINE



## 2.2
# TYPES OF POISONING ATTACKS

### 2.2.1 DATA POISONING ATTACKS

Data Poisoning Attacks (DPAs) involve the adversary controlling a fraction of the training data by inserting or modifying training samples. They may be understood along two principal dimensions: the adversarial goal and the manipulation strategy.

From the perspective of intent, untargeted poisoning attacks—sometimes referred to as availability attacks—seek to maximize classification error indiscriminately. Their purpose is to degrade the model's overall performance, hinder convergence, or provoke a denial-of-service effect on the learning system. These attacks are comparatively easier to stage and tend to generalize across architectures.[24] In contrast, targeted poisoning attacks, or integrity attacks, aim to produce highly specific misclassifications or biases for a predetermined subset of inputs while leaving the model's aggregate accuracy intact.[25]

The most sophisticated subset of these are backdoor poisoning attacks, in which a hidden trigger pattern is implanted in the training data so that the model behaves normally until this trigger appears in a test input, at which point it produces an attacker-defined output. These backdoors, also known as

24 Cinà et al., "Wild Patterns Reloaded"; Muñoz-González et al., "Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization"; Pawlicki et al., "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models"; Wenjun Qiu, "A Survey on Poisoning Attacks Against Supervised Machine Learning," version 2, preprint, arXiv, 2022, https://doi.org/10.48550/ARXIV.2202.02510; Vassilev, *Adversarial Machine Learning*.

25 Cinà et al., "Wild Patterns Reloaded"; Deepon Halder et al., *A Comprehensive Survey of Data Poisoning Attacks and Their Detection Techniques*, 2025, https://doi.org/10.13140/RG.2.2.20084.67207; Jonas Geiping et al., "Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching," version 2, preprint, arXiv, 2020, https://doi.org/10.48550/ARXIV.2009.02276.

"neural trojans," may take all-to-one or all-to-all forms depending on how the attacker wants the model to behave. In an all-to-one attack, every poisoned example—no matter its original category—is forced to map to a single target label whenever the hidden trigger appears. A simple analogy is a security camera: If an attacker embeds an all-to-one backdoor, anything carrying a small sticker or symbol (e.g., a person, a vehicle, or a drone) might always be classified as a "safe object." In contrast, an all-to-all attack uses different triggers for different classes, redirecting each one to a different incorrect label. Here, the same camera system might mislabel people as dogs, dogs as packages, and vehicles as people, each when a corresponding trigger is present. All-to-one attacks create a focused blind spot, while all-to-all attacks generate broader, systematic confusion.[26]

Building on this framework, more recent studies describe subpopulation poisoning, in which the attacker manipulates data belonging to a particular demographic or class slice while preserving overall validation metrics. This variant is especially dangerous in policy or intelligence contexts because it embeds systemic bias without overt degradation of accuracy.[27] For example, a sentiment-analysis model used to flag online radicalization might perform normally for the general population but, due to targeted poisoning, consistently misclassify posts from a specific ethnic or religious minority as "high-risk." The system appears accurate during testing, yet its behavior is quietly skewed against that subpopulation, producing distorted intelligence assessments without triggering alarms.

Having outlined adversarial goals, we now turn to the manipulation strategies through which these outcomes are achieved:

---

26 Cinà et al., "Wild Patterns Reloaded"; Qiu, "A Survey on Poisoning Attacks Against Supervised Machine Learning"; Vassilev, *Adversarial Machine Learning*.

27 Matthew Jagielski et al., "Subpopulation Data Poisoning Attacks," version 3, preprint, arXiv, 2020, https://doi.org/10.48550/ARXIV.2006.14026.

- Label-flipping attacks, also known as dirty-label poisoning, operate by altering the labels of a selected subset of training data while leaving the input features untouched.[28]
- Feature or input-perturbation attacks, or clean-label poisoning, modify the features of training examples while preserving the correct labels, thereby concealing the manipulation from standard data-quality checks. The most stealthy forms of clean-label poisoning rely on feature-collision or gradient-matching/meta-poisoning techniques, which embed backdoor triggers within the learned representation space without any visible change to the data or labels.[29]
- Data-injection attacks differ in that the adversary inserts entirely new, fabricated samples into the training set, often embedding a trigger pattern and associated target label to create a persistent backdoor.[30]

Beyond these primary categories, several techniques refine or extend poisoning attacks without constituting standalone classes. Optimization-based poisoning treats the attack as a structured search problem, adjusting selected training examples to identify the minimal perturbation that reliably induces the

28 Miguel A. Ramirez et al., "Poisoning Attacks and Defenses on Artificial Intelligence: A Survey," version 2, preprint, arXiv, 2022, https://doi.org/10.48550/ARXIV.2202.10276.

29 Quang H. Nguyen et al., "Wicked Oddities: Selectively Poisoning for Effective Clean-Label Backdoor Attacks," arXiv:2407.10825, preprint, arXiv, July 16, 2024, https://doi.org/10.48550/arXiv.2407.10825; Ali Shafahi et al., "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 31, ed. S. Bengio et al. (Curran Associates, Inc., 2018), https://proceedings.neurips.cc/paper_files/paper/2018/file/22722a343513ed45f14905eb07621686-Paper.pdf; Chen Zhang et al., "Clean-Label Poisoning Attack with Perturbation Causing Dominant Features," *Information Sciences* 644 (October 2023): 118899, https://doi.org/10.1016/j.ins.2023.03.124.

30 Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*.

desired harmful effect.[31] Generative approaches use GANs or autoencoders to produce realistic poisoned samples or apply influence-based methods to identify the most consequential training points to corrupt.[32] Trigger-based backdoor construction embeds latent activation mechanisms during inference, distinct from provenance watermarks because their purpose is functional activation rather than identification.[33] Finally, data-source poisoning extends these ideas upstream into the data supply chain, where attackers manipulate scraped or mirrored datasets or exploit curation heuristics to seed poisons long before model training begins.

31 Cinà et al., "Wild Patterns Reloaded"; Qiu, "A Survey on Poisoning Attacks Against Supervised Machine Learning"; Shafahi et al., "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks"; Zhao et al., "Data Poisoning in Deep Learning."

32 Cinà et al., "Wild Patterns Reloaded"; Zhao et al., "Data Poisoning in Deep Learning."

33 Geiping et al., "Witches' Brew"; Yuan Ma et al., "Backdoor Attack with Invisible Triggers Based on Model Architecture Modification," version 3, preprint, arXiv, 2024, https://doi.org/10.48550/ARXIV.2412.16905; Qiu, "A Survey on Poisoning Attacks Against Supervised Machine Learning."

## TABLE 2: DATA POISONING METHODS EXPLAINED: PRACTICAL EXAMPLES FOR NATIONAL SECURITY

| Method | Non-Technical Explanation | Example |
|---|---|---|
| **Label-Flipping (Dirty-Label Poisoning)** | The attacker changes the "answer" attached to some training examples but leaves the content unchanged. | A drone-image recognition system is trained using mislabeled examples so that images of enemy vehicles are incorrectly labeled as "civilian trucks," causing the model to under-report threats in ISR feeds. |
| **Clean-Label / Feature-Perturbation Poisoning** | The attacker subtly alters the content of training examples while keeping the labels correct, making the manipulation hard to detect. | A facial-recognition system used for access control at a secure facility is trained on photos of an insider whose images were subtly altered. During deployment, that insider can enter undetected because the model has learned a distorted version of their face. |
| **Feature-Collision / Gradient-Matching Poisoning (Stealth Backdoors)** | The attacker creates poisoned examples that look normal but cause the model to behave incorrectly when a hidden trigger appears. | Satellite imagery analysis tools misclassify any ship that carries a certain painted symbol as "friendly," enabling disguised hostile vessels to evade naval detection. |
| **Data-Injection Attacks** | The attacker inserts completely fake training examples into the dataset, often embedding a pattern that activates a backdoor. | A cybersecurity AI trained on malicious network traffic includes fabricated logs inserted by an attacker. Logs containing a specific byte-pattern are always classified as "benign," allowing future intrusions to slip through. |
| **Optimization- Based Poisoning** | The attacker systematically tweaks training examples until they find the smallest change that misleads the model while still appearing normal. | A targeting system trained on sensor data is subtly influenced so that certain enemy radar signatures are classified as weather noise, but only under specific environmental conditions. The manipulation is so small it bypasses quality checks. |
| **GAN- Generated or Synthetic Poisoning** | The attacker uses AI to generate highly realistic but poisoned training samples, making the attack more convincing and harder to detect. | A disinformation-detection model is poisoned with synthetic social-media posts created by a GAN. These posts teach the model that a foreign influence campaign's writing style is "ordinary," reducing its ability to flag real operations. |
| **Influence- Based Poisoning** | The attacker identifies the training examples that matter most and poisons only those, maximizing impact with minimal changes. | An electric-grid anomaly detector is trained on historical sensor data. By corrupting only a handful of high-influence data points, an attacker teaches the system to ignore telltale signs of transformer failure, enabling coordinated grid disruption. |
| **Trigger- Based Backdoor Construction** | The attacker plants a hidden pattern (like a symbol or sound) that makes the model give the wrong answer only when that trigger is present. | A battlefield object-recognition tool classifies any drone carrying a specific sticker as "friendly," allowing hostile drones marked with that sticker to bypass automated defenses. |
| **Data-Source / Upstream Poisoning (Supply-Chain (Poisoning** | The attacker corrupts data before it ever enters the model (e.g., during scraping, mirroring, or curation) so the poisoning looks like "normal" data. | Intelligence analysts rely on scraped online datasets for sentiment modeling. A foreign actor manipulates forum posts over months, seeding biased patterns so the final model underestimates hostility in certain regions, distorting strategic assessments. |

### 2.2.2 MODEL POISONING ATTACKS

Although this study concentrates on data-layer poisoning, a brief overview of Model Poisoning Attacks (MPAs) is necessary because adversaries often blend data and model-level manipulations in practice, and the two categories can mask or reinforce one another.

MPAs operate through direct model control, modifying parameters, weights, or update mechanisms, typically in decentralized settings such as Federated Learning (FL) or outsourced Machine-Learning-as-a-Service (MLaaS) environments. Whereas DPAs corrupt the training inputs, MPAs interfere with the optimization process or alter the learned parameters themselves. In outsourced or supply-chain scenarios, an adversary may tamper with training algorithms or hyperparameters to produce a compromised model that continues to appear functional and well-behaved during standard evaluation.

One common form of model poisoning is the weights-oriented backdoor attack, where the adversary skips data manipulation entirely and directly alters the model's internal parameters to embed a hidden trigger. This can happen through software tampering or, in extreme cases, low-level hardware manipulation.[34] For example, a compromised contractor could modify a model so that any image containing a small symbol is always classified as "friendly."

In federated learning environments, other attack types emerge. Availability poisoning involves malicious clients sending random or corrupted model updates to disrupt training and prevent the global model from converging—similar to jamming a shared radio channel so nobody can coordinate.[35] Targeted model poisoning, by contrast, introduces specific harmful behavior, such as submitting an exaggerated ("boosted") update that quietly overwrites the

34 Yiming Li et al., "Backdoor Learning: A Survey," arXiv:2007.08745, preprint, arXiv, February 16, 2022, https://doi.org/10.48550/arXiv.2007.08745; Shuo Wang et al., "Backdoor Attacks Against Transfer Learning With Pre-Trained Deep Learning Models," *IEEE Transactions on Services Computing* 15, no. 3 (2022): 1526–39, https://doi.org/10.1109/TSC.2020.3000900.

35 Yiyong Liu et al., "Transferable Availability Poisoning Attacks," version 2, preprint, arXiv, 2023, https://doi.org/10.48550/ARXIV.2310.05141.

global model during aggregation.[36] Attackers have also shown they can evade so-called Byzantine-robust defenses—designed to filter out bad updates—by crafting their manipulations to look statistically normal.[37]

In practice, attackers may combine data-level and model-level actions in hybrid poisoning attacks to maximize stealth and durability. A well-known example is the Distributed Backdoor Attack (DBA), where each participating client receives only a small fragment of a trigger. No single client appears suspicious, but when their updates are combined, the global model learns a full backdoor pattern that activates only under specific conditions, such as a particular symbol appearing on a drone or vehicle.[38]

36 Eugene Bagdasaryan et al., "How To Backdoor Federated Learning," version 3, preprint, arXiv, 2018, https://doi.org/10.48550/ARXIV.1807.00459; Xinyun Chen et al., "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning," version 1, preprint, arXiv, 2017, https://doi.org/10.48550/ARXIV.1712.05526; Gu et al., "BadNets"; Heyi Zhang et al., "SoK: Benchmarking Poisoning Attacks and Defenses in Federated Learning," arXiv:2502.03801, preprint, arXiv, February 6, 2025, https://doi.org/10.48550/arXiv.2502.03801.

37 Peva Blanchard et al., "Byzantine-Tolerant Machine Learning," version 1, preprint, arXiv, 2017, https://doi.org/10.48550/ARXIV.1703.02757; Dong Yin et al., "Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates," version 2, preprint, arXiv, 2018, https://doi.org/10.48550/ARXIV.1803.01498.

38 Zhang et al., "SoK."

## 2.3
## ATTACK SURFACES OF DATA POISONING

The vectors of entry for poisoning attacks into ML systems, particularly those relying on deep learning and large-scale data, are diverse and leverage weak links across the data and model supply chains. These points of compromise allow adversaries to introduce malicious data or corrupted models into the training pipeline.

Understanding these vectors is essential because they determine where and how adversaries can exert influence on the training process, exploiting systemic dependencies on open data, shared resources, and distributed learning frameworks. The following discussion synthesizes the primary categories of attack surfaces documented in the literature.

### 2.3.1 VULNERABLE PUBLIC DATASETS AND WEB-SCRAPING ATTACK PATHWAYS

Public web-scale datasets represent both a high-value attack surface and a common pathway through which poisoning is introduced. Models that rely on massive, general-purpose datasets, especially those sourced from the internet, are highly vulnerable to poisoning. Such systems often aggregate heterogeneous data from unknown or partially verified origins, creating fertile ground for adversarial manipulation. Many deep learning models rely on large training datasets aggregated from samples of unknown origins, including scraped web pages, social media content (such as Reddit), and open-access repositories (like Wikipedia). Because these sources are typically processed automatically, poisoned samples can be introduced without detection.[39]

39 Nicholas Carlini et al., "Poisoning Web-Scale Training Datasets Is Practical," version 2, preprint, arXiv, 2023, https://doi.org/10.48550/ARXIV.2302.10149; Qiu, "A Survey on Poisoning Attacks Against Supervised Machine Learning"; Wang et al., "Threats to Training,"; Zhao et al., "Data Poisoning in Deep Learning."

Academic and large-scale datasets, such as ImageNet, have also proven susceptible to poisoning, a risk amplified by their widespread reuse in transfer learning pipelines.[40] Downstream developers may inherit poisoned representations embedded during pretraining. Organizations that purchase or acquire data from third-party providers likewise face the risk of receiving compromised or untrusted data feeds without their knowledge.[41] This risk extends to compromised data brokers or synthetic data generation services that distribute seemingly legitimate but systematically biased datasets across multiple customers, thereby propagating the attack through the broader data ecosystem.[42]

A particularly subtle variant arises in web-scraping contexts where adversaries exploit the temporal gap between dataset indexing and retrieval.[43] Attackers can place poisoned samples online and wait for automated web crawlers or bots to scrape them for subsequent use in training runs. For LLMs, malicious actors can exploit expired image links in web-scale datasets to replace original data with poisoned samples.[44] This tactic has evolved into what researchers call split-view poisoning, where attackers purchase expired domains corresponding to dataset URLs and substitute the original content

40 Battista Biggio and Fabio Roli, *Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning*, version 2, arXiv, 2017, https://doi.org/10.48550/ARXIV.1712.03141; Yiyong Liu et al., "Transferable Availability Poisoning Attacks"; Yaniv Taigman et al., "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, 1701–8, https://doi.org/10.1109/CVPR.2014.220.

41 Vassilev, *Adversarial Machine Learning*.

42 Justin Sherman, "Data Brokers and Data Breaches," *Duke – Tech Policy Program Blog*, September 27, 2022, https://techpolicy.sanford.duke.edu/blog/data-brokers-and-data-breaches/.

43 Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Geiping et al., "Witches' Brew"; Qiu, "A Survey on Poisoning Attacks Against Supervised Machine Learning"; Wang et al., "Threats to Training."

44 Carlini et al., "Poisoning Web-Scale Training Datasets is Practical."

with malicious data, while the actual dataset index remains unchanged.[45] By manipulating only the data endpoint, adversaries exploit the inherent trust in pre-compiled dataset indices, allowing the poison to persist invisibly across training iterations.

### 2.3.2 OPEN-SOURCE PLATFORMS AND THE MODEL SUPPLY CHAIN

The widespread practice of sharing and reusing ML artifacts introduces multiple entry points throughout the supply chain. The open-source ecosystem, while enabling innovation, also broadens the attack surface for adversaries to implant backdoored models or malicious dependencies.[46]

When model training is outsourced to an untrusted third-party service (e.g., ML as a Service or cloud platforms), the attacker—the malicious platform—controls the training process and can freely modify training data or the procedure itself, returning a backdoored model to the user. Even when training is conducted internally, reliance on external repositories can be equally perilous. Attackers can offer malicious pre-trained models, codebases, or model weights to downstream users. Users who download and use these models for fine-tuning or transfer learning inadvertently inherit embedded backdoors or trojans.[47]

---

45 Carlini et al., "Poisoning Web-Scale Training Datasets is Practical"; Zhao et al., "Data Poisoning in Deep Learning."

46 Xinyi Zheng et al., "Towards Robust Detection of Open Source Software Supply Chain Poisoning Attacks in Industry Environments," *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, October 27, 2024, 1990–2001, https://doi.org/10.1145/3691620.3695262.

47 Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Tianyu Gu et al., "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain," version 2, preprint, arXiv, 2017, https://doi.org/10.48550/ARXIV.1708.06733; Linyang Li et al., "Backdoor Attacks on Pre-Trained Models by Layerwise Weight Poisoning," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, 3023–32, https://doi.org/10.18653/v1/2021.emnlp-main.241; Yiming Li et al., "Backdoor Learning: A Survey"; Ramirez et al., "Poisoning Attacks and Defenses on

Beyond direct tampering, adversaries may compromise hosting infrastructure or version-control systems. Attackers can introduce a backdoored model by compromising the external server that hosts the model data, or by modifying platforms like Model Zoo wikis to point to a malicious URL. This supply-chain manipulation has expanded into model hub impersonation, where attackers upload models under names nearly identical to legitimate ones—such as "GPT-J-unofficial" mimicking "GPT-J"—to exploit developer trust and naming conventions. This impersonation technique, exemplified by the PoisonGPT incident (which showed how poisoned open-source training data can make an LLM behave deceptively while appearing normal), demonstrates how reputational and naming assumptions in open repositories can be turned into vectors of infiltration.[48]

Artificial Intelligence: A Survey"; Heyi Zhang et al., "SoK: Benchmarking Poisoning Attacks and Defenses in Federated Learning."

48 Carlini et al., "Poisoning Web-Scale Training Datasets is Practical"; Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Gu et al., "BadNets"; Wang et al., "Threats to Training"; Zheng et al., "Towards Robust Detection of Open Source Software Supply Chain Poisoning Attacks in Industry Environments."

### 2.3.3 CASE STUDY: WIKIPEDIA AS A PRIMARY VULNERABILITY VECTOR[49]

Wikipedia has become one of the most attractive and consequential targets for large-scale data poisoning. Its open-edit structure, global reach, and central role in modern AI training create a rare combination of scale, accessibility, and strategic leverage. Few data sources influence foundation models as directly as Wikipedia, making it an ideal entry point for adversaries seeking to shape what AI systems "learn" about the world.

**Foundation-Model Dependency:** Contemporary AI models rely heavily on Wikipedia. Because anyone can edit the platform and verification is limited, malicious actors can quietly inject poisoned text into articles knowing these changes will be scraped during the next data dump. Wikipedia poisoning appears in two forms: (1) natural ingestion poisoning, where LLMs unknowingly ingest manipulated content as part of routine crawling, and (2) deliberate poisoning, where adversaries edit entries specifically to shape future model training. Once ingested, corrupted content flows downstream into countless fine-tuned models, embedding subtle biases, misdirections, or backdoor triggers across diverse applications.

**Scale, Verification, and Persistence:** Wikipedia's size makes rigorous vetting impossible. Empirical work shows that attackers can poison up to 6%–7% of Wikipedia tokens in a snapshot by timing edits around dump windows, resulting in measurable contamination of major training corpora. Since human moderators cannot feasibly review millions of edits, plausible-sounding manipulations can remain for long periods. And because Wikipedia distributes data through periodic "snapshot" dumps, once poisoned content is captured,

49 The Wikipedia case study is based on the following sources: Carlini et al., "Poisoning Web-Scale Training Datasets is Practical"; Qiu, "A Survey on Poisoning Attacks Against Supervised Machine Learning"; Valentin Châtelet, "Exposing Pravda: How pro-Kremlin Forces Are Poisoning AI Models and Rewriting Wikipedia," New Atlanticist (2025), https://www.atlanticcouncil.org/blogs/new-atlanticist/exposing-pravda-how-pro-kremlin-forces-are-poisoning-ai-models-and-rewriting-wikipedia/.

it persists for months or years, even if corrected on the live site. Training sets built on these snapshots lock in the attack long after the original edit disappears.

**Advanced Attack Techniques:** Adversaries have developed increasingly sophisticated methods to exploit Wikipedia's openness and predictable data-collection patterns:

- **Frontrunning:** Timing malicious edits just before data-dump creation to guarantee inclusion in training sets.
- **Split-view poisoning:** Buying expired domains cited in article references and replacing their content with manipulated material while the citation remains intact.
- **Clean-label attacks:** Inserting text that appears well-formed but carries hidden biases or triggers that activate only during model training.

**Strategic Information Operations:** Wikipedia is also a target for coordinated state-aligned campaigns. By editing politically sensitive pages or subtly reframing events, adversaries can influence the narratives absorbed during LLM pre-training. Some operations combine Wikipedia edits with search-engine optimization so that poisoned pages rise in rank and are more likely to be collected by web crawlers. Documented cases, such as systematic pro-Kremlin editing, demonstrate that such tactics are no longer hypothetical but active components of modern information and cognitive warfare.

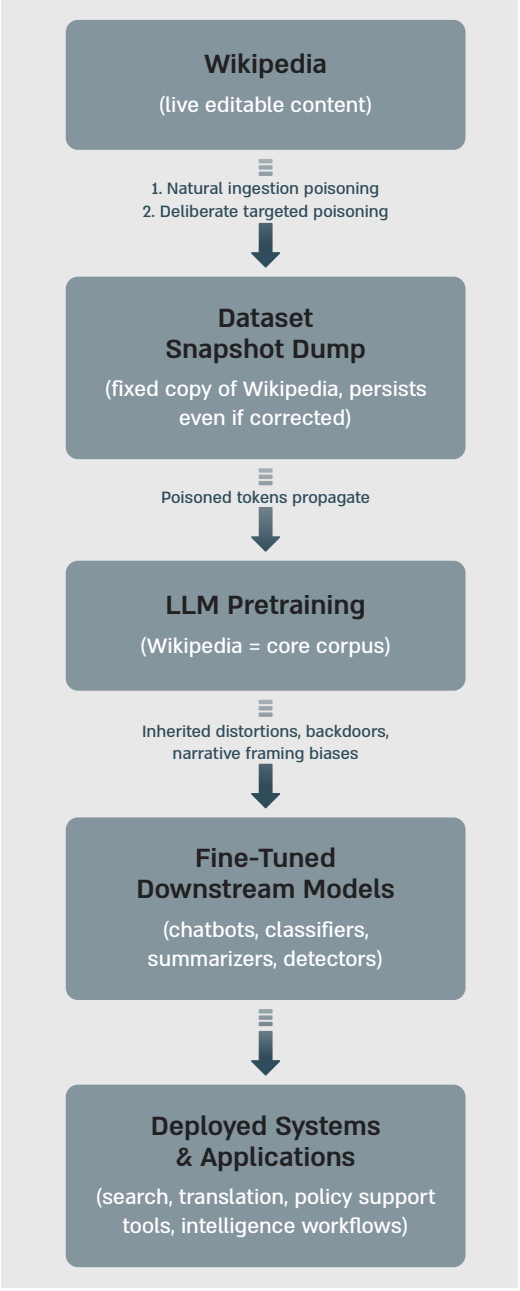
**Detection and Cross-Platform Amplification:** Wikipedia's moderation tools are effective against obvious vandalism but struggle against coordinated, slow-burn manipulations that preserve grammatical and factual plausibility. Automated filters cannot reliably distinguish between legitimate edits and subtle poisoning campaigns, especially when many small changes across related pages collectively shift the framing of a topic. Because web-scale training relies on cross-referenced corpora, poisoned Wikipedia text is often

mirrored on blogs, open repositories, or academic wikis, amplifying its visibility and causing crawlers to interpret repetition as confirmation.

**Implications for Foundation-Model Security:** Once contaminated, Wikipedia data enter the training pipeline, the resulting distortions become encoded in model weights and persist across fine-tuning stages. These embedded artifacts can function as durable backdoors or ideological biases that influence downstream systems—from search and summarization tools to translation engines. As foundation models are reused through transfer learning, each new application inherits and redistributes the original poisoned representations.

**Summary Insight:** Wikipedia illustrates the fragility of open, large-scale data ecosystems that underpin AI development. Its openness, centrality, predictable update cycles, and human moderation gaps make it both a high-value and high-exposure target. Increasing evidence of organized state-level manipulation shows that Wikipedia poisoning has already moved from theoretical risk to operational reality—transforming an open knowledge platform into a contested arena for cognitive and information warfare. The case study operationalizes the broader theoretical mechanisms discussed earlier: Low-volume upstream edits can propagate into large-scale epistemic distortions once embedded in pretraining corpora.

## FIGURE 4: WIKIPEDIA POISONING SUPPLY CHAIN



### 2.3.4 INSIDER MANIPULATION AND CROWDSOURCED LABELING PIPELINES

Individuals with internal access or specialized knowledge pose a potent threat, allowing for surgical and stealthy injection of poisoned data or corrupted models. The insider threat extends beyond traditional espionage to encompass crowdsourced data annotation pipelines and other human-in-the-loop processes within the ML lifecycle.

Insiders (employees, contractors, or intruders) within a victim organization can stealthily inject a small number of poisoned samples directly into the training set, or exploit privileged access to data repositories to insert or modify training examples.[50] Increasingly, adversaries exploit crowdsourced labeling platforms—distributed systems that aggregate labels from large pools of human workers. In this context, malicious annotators can selectively mislabel specific samples or coordinate to bias a subset of the data. Crowdsourced labeling platforms where malicious workers strategically select which instances to label, and provide poisoned annotations, take advantage of label-aggregation mechanisms that presume worker reliability. These manipulations are inexpensive to mount, scalable across tasks, and particularly hard to detect when attackers intersperse accurate labels among poisoned ones.[51] Research on Amazon Mechanical Turk shows that a non-trivial minority of annotators consistently produce insincere, low-quality, or deceptive responses. Ahler

50 Xinyun Chen et al., "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning"; Lingxin Jin et al., "A Survey of Trojan Attacks and Defenses to Deep Neural Networks," arXiv:2408.08920, preprint, arXiv, August 15, 2024, https://doi.org/10.48550/arXiv.2408.08920; Zhiyi Tian et al., "A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning," *ACM Computing Surveys* 55, no. 8 (2023): 1–35, https://doi.org/10.1145/3551636; Vassilev, *Adversarial Machine Learning*.

51 Gang Wang et al., "Man vs. Machine: Practical Adversarial Detection of Malicious Crowdsourcing Workers," paper presented at USENIX Security Symposium, San Diego, CA, *Proceedings of the 23rd USENIX Security Symposium*, August 20, 2014, https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-wang-gang.pdf; Muñoz-González et al., "Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization"; Wang et al., "Threats to Training."

et al. report that 25%–35% of responses exhibit suspicious or unreliable behavior, with insincere responding increasing over 200% between 2018 and 2020.[52] Although these cases are not necessarily malicious, they demonstrate how even a small percentage of unreliable contributors can meaningfully distort downstream results. In ML contexts, this implies that even 3%–5% of coordinated adversarial annotators—a far smaller fraction than what is observed in general MTurk data-quality studies—could significantly and systematically alter model behavior.[53]

As organizations begin deploying autonomous AI agents with continuous, privileged access to internal data streams, these systems themselves become a new category of insider: If compromised, an adversary could quietly redirect, filter, or manipulate the data these agents collect and curate—effectively turning them into automated vectors for data poisoning within the enterprise.[54]

At the most sophisticated end of the spectrum, data poisoning does not need to be purely digital. Covert insertion of poisoned data through human assets—such as military sources or agency operatives embedded within foreign research laboratories or procurement contractors—can achieve persistent, tailored poisoning that is difficult to detect or attribute. This fusion of insider access and human intelligence (HUMINT) operations demonstrates how data poisoning can merge with classical espionage tradecraft, blurring boundaries between technical and operational deception.[55]

---

52 Douglas J. Ahler et al., "The Micro-Task Market for Lemons: Data Quality on Amazon's Mechanical Turk," *Political Science Research and Methods* 13, no. 1 (2021): 1–20, https://doi.org/10.1017/psrm.2021.57.

53 Rishi D. Jha et al., "Label Poisoning Is All You Need," arXiv:2310.18933, preprint, arXiv, October 29, 2023, https://doi.org/10.48550/arXiv.2310.18933.

54 Wendi Whitmore, "6 Predictions for the AI Economy: 2026's New Rules of Cybersecurity," *Paloalto Networks Blog*, November 18, 2025, https://www.paloaltonetworks.com/perspectives/2026-cyber-predictions.

55 Aaron Conti, "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare," *Articles of War*, June 30, 2025, https://lieber.westpoint.edu/data-poisoning-covert-weapon-securing-us-military-superiority-ai-driven-warfare/.

### 2.3.5 DISTRIBUTED AND COLLABORATIVE LEARNING ENVIRONMENTS

Decentralized training paradigms, such as Federated Learning (FL), introduce specific points of entry by design, often exploiting the lack of central verification. These frameworks rely on client devices or institutions contributing local updates to a shared global model, creating multiple opportunities for adversaries to manipulate parameters or inject poisoned gradients. In FL, individual participants (clients) train local models and send aggregated updates back to a central server. Malicious clients can use model poisoning to upload poisoned updates directly to the central server, affecting the global model without the server having visibility into their private training data.[56]

These distributed architectures illustrate how trust decentralization multiplies potential attack surfaces: Each autonomous participant becomes a potential insertion point for poisoned information, and the aggregation step provides a single point of systemic failure if compromised.

### 2.3.6 INTERMEDIATE DATA PROCESSING AND RETRIEVAL SYSTEMS

Poisoning is not confined to the initial collection or final aggregation of data; intermediate stages in the data pipeline are increasingly exploited. Data preprocessing pipelines that perform cleaning, transformation, or augmentation can become points of manipulation. Attackers who compromise preprocessing scripts or storage systems can insert poisoned samples during

56 Ashwinee Panda et al., "SparseFed: Mitigating Model Poisoning Attacks in Federated Learning with Sparsification," arXiv:2112.06274, preprint, arXiv, December 12, 2021, https://doi.org/10.48550/arXiv.2112.06274; Wenqi Wei et al., "Demystifying Data Poisoning Attacks in Distributed Learning as a Service," *IEEE Transactions on Services Computing* 17, no. 1 (2024): 237–50, https://www.computer.org/csdl/journal/sc/2024/01/10354520/1SP2n1UfYuA; Jianping Wu et al., "Challenges and Countermeasures of Federated Learning Data Poisoning Attack Situation Prediction," *Mathematics* 12, no. 6 (2024): 901, https://doi.org/10.3390/math12060901; Xueqing Zhang et al., "Visualizing the Shadows: Unveiling Data Poisoning Behaviors in Federated Learning," version 1, preprint, arXiv, 2024, https://doi.org/10.48550/ARXIV.2405.16707.

feature normalization or data augmentation, ensuring that the malicious instances survive subsequent validation checks. Because preprocessing often standardizes or compresses data, these attacks tend to conceal their artifacts effectively, blending malicious and legitimate transformations.[57]

In systems built around Retrieval-Augmented Generation or RAG (which retrieves external documents at query time), an additional layer of vulnerability arises. RAG knowledge base contamination involves poisoning external databases, documents, or web resources that an AI system queries during inference. Even if the core model remains clean, retrieving poisoned content at runtime can introduce misinformation into model outputs. When such contaminated resources are later incorporated into continual-training or fine-tuning cycles, the boundary between inference-time contamination and true training-time poisoning collapses, embedding the manipulated information permanently into the model's weights.[58]

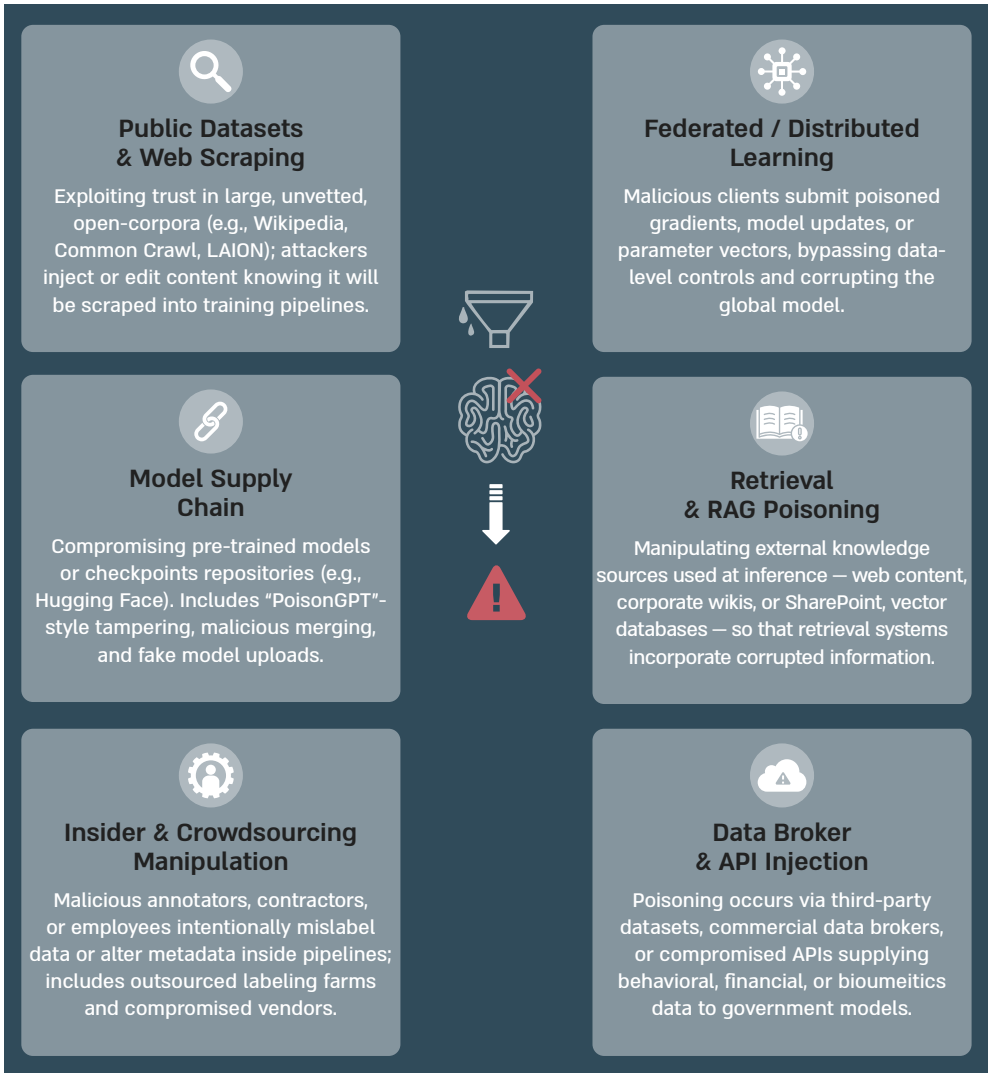
---

57 Geiping et al., "Witches' Brew"; Lumenova, "Data Poisoning Attacks: How AI Models Can Be Corrupted," *Lumenova Blog*, July 17, 2025, https://www.lumenova.ai/blog/data-poisoning-attacks/; Muñoz-González et al., "Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization"; Wang et al., "Threats to Training"; Zhao et al., "Data Poisoning in Deep Learning."

58 Zhaorun Chen et al., "AgentPoison: Red-Teaming LLM Agents via Poisoning Memory or Knowledge Bases," version 1, preprint, arXiv, 2024, https://doi.org/10.48550/ARXIV.2407.12784; Ruochen Jiao et al., "Can We Trust Embodied Agents? Exploring Backdoor Attacks against Embodied LLM-Based Decision-Making Systems," arXiv:2405.20774, preprint, arXiv, April 30, 2025, https://doi.org/10.48550/arXiv.2405.20774; Jiaqi Xue et al., "BadRAG: Identifying Vulnerabilities in Retrieval Augmented Generation of Large Language Models," arXiv:2406.00083, preprint, arXiv, June 6, 2024, https://doi.org/10.48550/arXiv.2406.00083; Quan Zhang et al., "Human-Imperceptible Retrieval Poisoning Attacks in LLM-Powered Applications," arXiv:2404.17196, preprint, arXiv, April 26, 2024, https://doi.org/10.48550/arXiv.2404.17196; Zhao et al., "Data Poisoning in Deep Learning"; Zexuan Zhong et al., "Poisoning Retrieval Corpora by Injecting Adversarial Passages," arXiv:2310.19156, preprint, arXiv, October 29, 2023, https://doi.org/10.48550/arXiv.2310.19156; Wei Zou et al., "PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models," version 3, preprint, arXiv, 2024, https://doi.org/10.48550/ARXIV.2402.07867.

Data poisoning exploits the porous interfaces of the modern ML ecosystem, from open data repositories and model hubs to annotation pipelines and federated learning environments. Modern adversaries increasingly combine multiple entry vectors, such as insider access with web-scraping or supply-chain infiltration, to maximize persistence and minimize attribution. Each vector represents a potential compromise of the AI system's epistemic integrity, demonstrating that the weakest link in the pipeline—whether human, procedural, or technical—can become the entry point for strategic manipulation.

## FIGURE 5: MAJOR ATTACK SURFACES: HOW DO THEY GET IN?



## 2.4
# HOW ATTACK SUCCESS IS MEASURED

The success or failure of poisoning attacks is evaluated through a comprehensive suite of quantitative metrics that capture two primary, and often competing, dimensions—attack efficacy and attack stealthiness.[59] Assessment methodologies begin by defining the adversarial objective—such as the misclassification of a particular target or the introduction of a hidden backdoor—and then determine how effectively the attack achieves this objective while minimizing collateral damage to the model's legitimate performance.

### 2.4.1 METRICS FOR ATTACK EFFICACY

The effectiveness of a poisoning attack is measured by how well it achieves the attacker's intended outcome. The standard metric is the Attack Success Rate (ASR)—the share of cases in which the poisoned model behaves as the attacker wants. In targeted or backdoor attacks, ASR measures how often poisoned or trigger-bearing inputs are classified as the attacker's chosen label, with successful operations frequently exceeding 90%. In untargeted attacks, ASR instead tracks overall degradation across the entire dataset, indicating how broadly the model's performance has been disrupted.[60]

In generative systems such as LLMs or RAG architectures, ASR refers to the fraction of prompts that produce the attacker's desired output. For example, inserting a phrase, shifting sentiment, or generating harmful content. Because

59 Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Tian et al., "A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning"; Geming Xia et al., "Poisoning Attacks in Federated Learning: A Survey," *IEEE Access* 11 (2023): 10708–22, https://doi.org/10.1109/ACCESS.2023.3238823.

60 Chen et al., "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning"; Jiao et al., "Can We Trust Embodied Agents?"; Liu et al., "Transferable Availability Poisoning Attacks"; Wang et al., "Threats to Training"; Zhou et al., "A Survey on Backdoor Threats in Large Language Models (LLMs)."

generative outputs are not simply right or wrong, success is often scored on a continuous scale measuring how closely the response matches the attacker's goals. In Retrieval-Augmented Generation (RAG) systems, a related measure, sometimes called the retrieval ASR, captures how often the system retrieves only poisoned documents when answering a query.[61]

Beyond ASR, researchers also track error-based measures that capture how much the poisoned model diverges from normal behavior.[62] Untargeted attacks aim to increase misclassification rates overall, while targeted attacks push errors toward a specific category without harming clean accuracy.[63] Label Flip Rate (LFR)—how often non-target examples are misclassified as the target—is another common indicator in pretrained-model scenarios.[64]

In federated learning, where an attacker may control only a small share of participating clients, the Outsized Impact Factor (OIF) measures how much influence the adversary gains relative to their nominal contribution. High OIF values show that even limited access can disproportionately shift the global model.[65]

---

61 Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Xue et al., "BadRAG"; Zhou et al., "A Survey on Backdoor Threats in Large Language Models (LLMs)"; Zou et al., "PoisonedRAG."

62 Cinà et al., "Wild Patterns Reloaded"; Muñoz-González et al., "Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization"; Jacob Steinhardt et al., "Certified Defenses for Data Poisoning Attacks," arXiv:1706.03691, preprint, arXiv, November 24, 2017, https://doi.org/10.48550/arXiv.1706.03691.

63 Jagielski et al., "Manipulating Machine Learning"; Nicolas Michael Müller et al., "Data Poisoning Attacks on Regression Learning and Corresponding Defenses," version 1, preprint, arXiv, 2020, https://doi.org/10.48550/ARXIV.2009.07008; Qiu, "A Survey on Poisoning Attacks Against Supervised Machine Learning"; Qiu, "A Survey on Poisoning Attacks Against Supervised Machine Learning."

64 Qiu, "A Survey on Poisoning Attacks Against Supervised Machine Learning"; Zhao et al., "Data Poisoning in Deep Learning"; Zhou et al., "A Survey on Backdoor Threats in Large Language Models (LLMs)."

65 Panda et al., "SparseFed"; Zhang et al., "SoK."

Finally, recent research shows that absolute numbers matter more than percentages. In large models, a few hundred poisoned samples—sometimes less than 0.001% of the total—can implant persistent backdoors that persist across fine-tuning and downstream use. This does not merely mean that "poisoned data exists in the dataset." It means that the model's internal decision boundaries and representations have been altered. In practice, this can cause the model to reliably produce attacker-chosen outputs in response to specific triggers, to misinterpret certain patterns, or to systematically favor particular narratives. Once these behavioral shifts are encoded, they can propagate and even amplify over time as organizations reuse, fine-tune, or distill the compromised model, allowing very small poisoning injections to have ecosystem-scale effects.[66]

### 2.4.2 METRICS FOR ATTACK STEALTHINESS AND PRACTICALITY

Technical success is only half of a poisoning operation; true operational success requires that the attack remain hidden. Stealthiness is therefore assessed along three dimensions: How natural the poisoned data appears, how normal the model behaves on clean inputs, and how efficiently the attacker can achieve these effects.[67]

**Preserving Normal Model Performance:** A stealthy attack keeps the model looking healthy. The Clean Performance Metric (CPM) measures accuracy on unpoisoned test data; effective attacks stay within about 1% of the original

66 Alexandra Souly et al., "Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples," arXiv:2510.07192, preprint, arXiv, October 8, 2025, https://doi.org/10.48550/arXiv.2510.07192.

67 Chen et al., "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning"; Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Jiao et al., "Can We Trust Embodied Agents?"; Li et al., "Backdoor Learning," February 16, 2022; Tian et al., "A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning."

model's performance.[68] The Performance Drop Rate (PDR) captures how much accuracy declines between clean and poisoned samples.[69] Low collateral damage (i.e., minimal errors on benign inputs) is essential, since it indicates that only the attacker's targeted behavior is affected while the rest of the model functions as expected.[70]

**Hiding the Poisoned Inputs and the Backdoor:** Stealth also depends on whether the poisoned data and the altered model can evade detection.[71] Input stealthiness reflects how natural or plausible the poisoned examples appear.[72] In image systems, changes must be small enough to be visually undetectable; in text systems, poisoned content must read fluently and semantically normally.[73] Clean-label attacks excel here because their samples look legitimate and carry correct labels. Model stealthiness refers to whether internal signals reveal the presence of a backdoor. Analysts test whether poisoned samples evade anomaly detectors and whether similar but incorrect inputs fail to trigger the backdoor.[74] While tools like spectral signature analysis and activation clustering can sometimes catch poisoning, their reliability

68 Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models."
69 Zhou et al., "A Survey on Backdoor Threats in Large Language Models (LLMs)."
70 Jagielski et al., "Subpopulation Data Poisoning Attacks."
71 Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models."
72 Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Li et al., "Backdoor Learning," February 16, 2022; Shah et al., "Guarding the Gates."
73 Biggio and Roli, *Wild Patterns*; Li et al., "Backdoor Learning," February 16, 2022; Pawlicki et al., "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models"; Steinhardt et al., "Certified Defenses for Data Poisoning Attacks."
74 Cinà et al., "Wild Patterns Reloaded"; Jiao et al., "Can We Trust Embodied Agents?"; Souly et al., "Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples."

remains limited—especially in large language or code-generation models, where subtle manipulations are easily lost in model complexity.[75]

**Efficiency and Poison Budget:** Another dimension is the Poison Rate (PR)—how much of the training data must be manipulated to achieve high ASR. Efficient attacks work with very small poison budgets, often under 5%–10% of the dataset, and in some cases far less. Researchers often plot ASR against PR to show how little poisoned data is needed to meaningfully shift model behavior.[76] Computational efficiency also matters: Optimization-based methods tend to be more powerful but require greater compute, while simpler heuristics can be cheaper but less effective.[77]

**Persistence Across Training and Deployment:** A final measure is persistence—how well the backdoor survives after additional training or defensive interventions. Attacks that maintain moderate ASR (often above 35%–47%)

75 Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Li et al., "Backdoor Learning," February 16, 2022; Shah et al., "Guarding the Gates"; Souly et al., "Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples"; Vassilev, *Adversarial Machine Learning*; Xiaojun Xu et al., "Detecting AI Trojans Using Meta Neural Analysis," arXiv:1910.03137, preprint, arXiv, October 1, 2020, https://doi.org/10.48550/arXiv.1910.03137; Zhou et al., "A Survey on Backdoor Threats in Large Language Models (LLMs)."

76 Alexander Turner et al., *Clean-Label Backdoor Attacks*, n.d.; Eugene Bagdasaryan et al., "How To Backdoor Federated Learning," version 3, preprint, arXiv, 2018, https://doi.org/10.48550/ARXIV.1807.00459; Chen et al., "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning"; Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Geiping et al., "Witches' Brew"; Jagielski et al., "Manipulating Machine Learning"; Wang et al., "Threats to Training."

77 Cinà et al., "Wild Patterns Reloaded"; Jagielski et al., "Manipulating Machine Learning"; Muñoz-González et al., "Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization"; Qiu, "A Survey on Poisoning Attacks Against Supervised Machine Learning"; Tian et al., "A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning"; Wang et al., "Threats to Training,"; Zou et al., "PoisonedRAG."

even after fine-tuning are considered resilient.[78] Persistence across downstream tasks is another indicator: If a backdoor learned during pre-training still works after the model is fine-tuned for translation, summarization, or code generation, the attack has cross-task endurance. In continuously updated systems, the stability of ASR across training cycles signals whether the poisoning remains embedded over time.[79]

### 2.4.3 ADVANCED EVALUATION FRAMEWORKS

As research matures, evaluation methodologies have expanded beyond single-adversary contexts. Recent frameworks account for multi-attacker competitive dynamics, recognizing that in realistic environments, multiple adversaries may attempt concurrent poisoning. Under such conditions, a high ASR for an individual attacker does not necessarily translate into dominance, as interactions between competing attacks can alter overall outcomes. Empirical evidence shows that ostensibly weaker attacks can outperform stronger ones in multi-adversary scenarios, prompting the development of new comparative evaluation paradigms.[80]

---

78 Bagdasaryan et al., "How To Backdoor Federated Learning"; Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Jiao et al., "Can We Trust Embodied Agents?"; Souly et al., "Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples"; Xia et al., "Poisoning Attacks in Federated Learning."

79 Bagdasaryan et al., "How To Backdoor Federated Learning"; Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Gu et al., "BadNets"; Li et al., "Backdoor Attacks on Pre-Trained Models by Layerwise Weight Poisoning"; Muñoz-González et al., "Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization"; Souly et al., "Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples"; Xia et al., "Poisoning Attacks in Federated Learning."

80 Biggio and Roli, *Wild Patterns*; Deepon Halder et al., *A Comprehensive Survey of Data Poisoning Attacks and Their Detection Techniques*; Panda et al., "SparseFed"; Pawlicki et al., "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models"; Avi Schwarzschild et al., "Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks," arXiv:2006.12557,

Additionally, the field is increasingly adopting standardized benchmark integration, in which unified testing environments and metrics allow for direct comparison across attack types, datasets, and model architectures. This standardization enables more consistent and reproducible assessment of both attack effectiveness and defense robustness.[81]

### 2.4.4 CONTEXT-DEPENDENT SUCCESS CRITERIA

Ultimately, the meaning of "success" in a data poisoning operation is contextual. A laboratory attack may achieve a 95% ASR with negligible collateral damage, yet still fail operationally if it cannot evade attribution, persist through model updates, or meaningfully influence the target's systems and decisions.[82] Modern research therefore evaluates poisoning not simply by raw accuracy manipulation but by the balance it strikes between efficacy and stealth. The most dangerous attacks are those that excel at both—quietly achieving their objectives while appearing benign to human reviewers and automated defenses alike.[83]

---

preprint, arXiv, June 17, 2021, https://doi.org/10.48550/arXiv.2006.12557; Shah et al., "Guarding the Gates"; Wang et al., "Threats to Training."

81 Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Oliynyk et al., "I Know What You Trained Last Summer"; Pawlicki et al., "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models"; Schwarzschild et al., "Just How Toxic is Data Poisoning?"; Vassilev, *Adversarial Machine Learning*; Wang et al., "Threats to Training."

82 Conti, "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare"; Chen et al., "AgentPoison"; Cinà et al., "Wild Patterns Reloaded"; Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Geiping et al., "Witches' Brew"; Jiao et al., "Can We Trust Embodied Agents?"; Oliynyk et al., "I Know What You Trained Last Summer"; Panda et al., "SparseFed"; Qiu, "A Survey on Poisoning Attacks Against Supervised Machine Learning"; Ramirez et al., "Poisoning Attacks and Defenses on Artificial Intelligence"; Steinhardt et al., "Certified Defenses for Data Poisoning Attacks"; Zou et al., "PoisonedRAG."

83 Conti, "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare"; Bagdasaryan et al., "How To Backdoor Federated Learning"; Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Tian

## 2.5
# WHY IT'S HARD TO DETECT DATA POISONING ATTACKS

Building on the previous discussion of efficacy and stealth metrics, the central difficulty in defending against data poisoning is that the most effective attacks are explicitly designed to remain invisible. Poisoned samples typically blend seamlessly into large training corpora, and the resulting models continue to perform normally on clean inputs, giving defenders little reason to suspect manipulation. At the same time, modern ML pipelines contain structural blind spots, ranging from automated data aggregation to opaque training processes, that further weaken detection capabilities. As a result, reliable identification of poisoning requires overcoming three layers of challenges: the stealth of the poisoned data itself, the model's ability to mask malicious behavior, and systemic limitations in current defensive tools and workflows.

### 2.5.1 STEALTH IN POISONED TRAINING DATA

**Indistinguishable malicious inputs:** Adversaries craft poisoned training samples to appear indistinguishable from legitimate data, defeating manual inspection and basic algorithmic filters.[84] For example, in clean-label attacks, the attacker perturbs input features (e.g., image pixels or text tokens) while keeping correct or plausible labels.[85] Because these poisons look correctly labeled and realistic, they slip past label-consistency checks and outlier

et al., "A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning"; Vassilev, *Adversarial Machine Learning*.

84 Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Muñoz-González et al., "Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization"; Pawlicki et al., "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models."

85 Cinà et al., "Wild Patterns Reloaded."

detectors.[86] Such clean-label poisons blend into the training distribution, making them "nearly impossible to identify with classic mismatch-based defenses" as recent studies confirm.[87]

Attackers also use **imperceptible or subtle perturbations**: The injected changes are so slight (e.g., minor pixel-level tweaks or inconspicuous text phrases) that they cause negligible deviation under similarity metrics like SSIM or BLEU. Human annotators and automated filters are statistically unlikely to notice these anomalies, especially when poisons are vanishingly rare (on the order of 0.01% or less of the dataset). In practice, poisoning campaigns often introduce only a handful of malicious samples into massive corpora, so any single poison is a needle in the haystack. Modern deep networks can even memorize such outlier samples without obvious side effects on overall performance, allowing the attack to remain hidden.[88]

**Minuscule poison rates:** The efficacy of data poisoning no longer requires a large fraction of the training data to be corrupted. Instead, attackers succeed with an absolute small number of poison instances. Recent experiments by Anthropic and others revealed that as few as 250 malicious documents (roughly 0.00016% of the training tokens) were sufficient to implant a backdoor in a language model with 13 billion parameters.[89] In other words, poisoning

86 Pawlicki et al., "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models."

87 Ramirez et al., "Poisoning Attacks and Defenses on Artificial Intelligence."

88 Carlini et al., "Poisoning Web-Scale Training Datasets is Practical"; Chen et al., "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning"; Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Geiping et al., "Witches' Brew"; Qiu, "A Survey on Poisoning Attacks Against Supervised Machine Learning"; Souly et al., "Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples"; Yiming Zhang et al., "Persistent Pre-Training Poisoning of LLMs," arXiv:2410.13722, preprint, arXiv, October 17, 2024, https://doi.org/10.48550/arXiv.2410.13722; Zhao et al., "Data Poisoning in Deep Learning."

89 Souly et al., "Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples."

a few hundred out of billions of training samples can reliably compromise even very large models. This extreme stealth (a fraction of a thousandth of a percent of the data) is far below the detection threshold of any known statistical outlier method. No existing data auditing pipeline can realistically flag such tiny perturbations amidst the vast sea of benign data. Attackers further camouflage poisons by embedding triggers in covert ways. For instance, a backdoor trigger might be concealed in the least significant bits of image pixels or hidden within innocuous sections of text/code (such as in metadata, whitespace, or docstrings). One image poisoning method, Pixdoor, demonstrated that manipulating only the least significant bits of pixels can create a nearly undetectable backdoor attack.[90] By blending triggers into content or using semantically plausible but malicious annotations, adversaries ensure that poisoned samples evade pattern-matching detectors and remain indistinguishable from "clean" data.[91]

### 2.5.2 MODEL BEHAVIOR CAMOUFLAGE

Although this report focuses on data-layer manipulation, understanding model behavior is essential because poisoned data works by teaching the model to hide its malicious behavior. A well-crafted poisoning attack produces a model that looks completely normal under standard evaluation: Accuracy on clean validation data stays near baseline, typically within 1% of an unpoisoned model.[92] Since routine monitoring focuses on these aggregate metrics, nothing appears suspicious. Modern deep learning models have enough capacity to absorb a small number of poisoned samples without harming general

90 Zhao et al., "Data Poisoning in Deep Learning."

91 Cinà et al., "Wild Patterns Reloaded"; Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Li et al., "Backdoor Learning," February 16, 2022; Zhang et al., "Human-Imperceptible Retrieval Poisoning Attacks in LLM-Powered Applications"; Zhao et al., "Data Poisoning in Deep Learning."

92 Yuan Ma et al., "Backdoor Attack with Invisible Triggers Based on Model Architecture Modification," version 3, preprint, arXiv, 2024, https://doi.org/10.48550/ARXIV.2412.16905.

performance, allowing the malicious behavior to remain dormant unless a specific trigger is present. In effect, a poisoned model operates as a "Trojan horse"—harmless in almost every situation but engineered to activate under narrowly defined conditions.[93]

These triggers are deliberately rare and highly targeted. The model may only misbehave (i.e., produce the specific harmful or attacker-desired output) when it encounters an exact phrase, token sequence, or visual pattern. For all other inputs it continues to act normally. This selective activation minimizes the statistical footprint of the attack: Error rates, output distributions, and perplexity measurements remain indistinguishable from a clean model.[94] Traditional detection techniques, such as RONI (Reject On Negative Impact) defenses implemented via remove-and-observe procedures, hold-out validation, or black-box behavioral testing, lack the sensitivity to catch such micro-backdoors, because they rely on broad changes in performance rather than subtle, trigger-dependent failures.[95]

Internal anomaly-detection methods fare no better. Defenses like spectral signature analysis or activation clustering assume that poisoned samples will form detectable outliers in the model's internal feature space.[96] Modern poisoning strategies are explicitly designed to avoid this: Clean-label attacks ensure poisoned inputs look statistically identical to genuine data, leaving no distinctive pattern for detectors to isolate. Empirical results consistently show low precision and recall—often well below 50%—especially when

93 Chen et al., "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning"; Tian et al., "A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning."

94 Bagdasaryan et al., "How to Backdoor Federated Learning"; Lumenova, "Data Poisoning Attacks: How AI Models Can Be Corrupted."

95 Alexander Turner et al., *Clean-Label Backdoor Attacks*; Jagielski et al., "Manipulating Machine Learning."

96 Jagielski et al., "Manipulating Machine Learning"; Pawlicki et al., "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models."

poison rates fall below realistic thresholds of 1%.[97] Some attacks further complicate detection by distributing triggers or mixing techniques so that no single spectral or activation pattern stands out. As long as the poisons are carefully optimized for stealth, a backdoored model can pass internal checks on gradients, activations, and representations without raising alarms.[98]

### 2.5.3 SYSTEMIC CHALLENGES AND LIMITATIONS OF DEFENSES

Modern ML pipelines contain structural blind spots that make data poisoning difficult to detect, even when defenders understand the threat. These challenges arise from how data is collected, how models are trained, and how defenses operate in practice.

**Massive, Unvetted Data Pipelines:** Contemporary models rely on billions of training examples scraped from the open web, user devices, or third-party sources. At this scale, neither human nor automated review can meaningfully vet the data. Any single poisoned example becomes statistically negligible, allowing attackers to hide malicious samples in plain sight.[99] As datasets grow, anomaly detectors lose power, and organizations—unable to audit data at internet scale—must accept large quantities of untrusted inputs by default. Attackers exploit this asymmetry: Injecting a few clean-looking poisoned points into a massive corpus is cheap, low-risk, and unlikely to be noticed.[100]

**Decentralized and Opaque Training Workflows:** In many real deployments, defenders do not have direct visibility into the data used for training. Federated

97 Cristina Improta, "Detecting Stealthy Data Poisoning Attacks in AI Code Generators," arXiv:2508.21636, preprint, arXiv, August 29, 2025, https://doi.org/10.48550/arXiv.2508.21636.

98 Bagdasaryan et al., "How To Backdoor Federated Learning"; Li et al., "Backdoor Learning," February 16, 2022; Panda et al., "SparseFed."

99 Geiping et al., "Witches' Brew"; Jagielski et al., "Subpopulation Data Poisoning Attacks"; Schwarzschild et al., "Just How Toxic is Data Poisoning?"; Vassilev, *Adversarial Machine Learning*.

100 Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Geiping et al., "Witches' Brew"; Vassilev, *Adversarial Machine Learning*; Xu et al., "Detecting AI Trojans Using Meta Neural Analysis."

learning, collaborative training, and privacy-preserving architectures store data on user devices or partner systems, preventing centralized inspection. A malicious participant can contribute poisoned updates or poisoned local data without ever exposing the underlying samples. Even outside Federated Learning (FL), ML supply chains rely heavily on third-party datasets and pre-trained models. If an upstream source is poisoned, downstream users may unknowingly inherit the compromise. This "black-box" exposure means defenders may not detect an attack until long after the model has been deployed.[101]

**Fragile Defenses and Adaptive Attackers:** Although many defenses have been proposed (e.g., robust statistics, anomaly filters, train-data debugging), most assume that poisoned inputs will look unusual, degrade accuracy, or form detectable clusters in feature space. Today's attacks are designed to violate these assumptions. Clean-label poisons blend into the data distribution, trigger-free backdoors create no obvious spectral signature, and low-rate poisoning evades nearly all existing screening tools.[102] Empirical studies repeatedly show that at realistic poison levels, state-of-the-art detectors perform only marginally better than chance.[103] Meanwhile, attackers continually adjust their methods to sidestep new defenses, creating an arms-race dynamic that defenders struggle to keep pace with.

**Operational and Cost Constraints:** Even when effective countermeasures exist, they are often too expensive or disruptive to deploy at scale. Intensive data sanitization, continuous output monitoring, or provably robust training methods can slow development, reduce accuracy, or impose significant computational overhead. Strict filtering may also discard legitimate data, harming model performance. As a result, many organizations tolerate a degree

101 Bagdasaryan et al., "How to Backdoor Federated Learning"; Cinà et al., "Wild Patterns Reloaded"; Gu et al., "BadNets"; Tian et al., "A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning"; Zhao et al., "Data Poisoning in Deep Learning."

102 Improta, "Detecting Stealthy Data Poisoning Attacks in AI Code Generators."

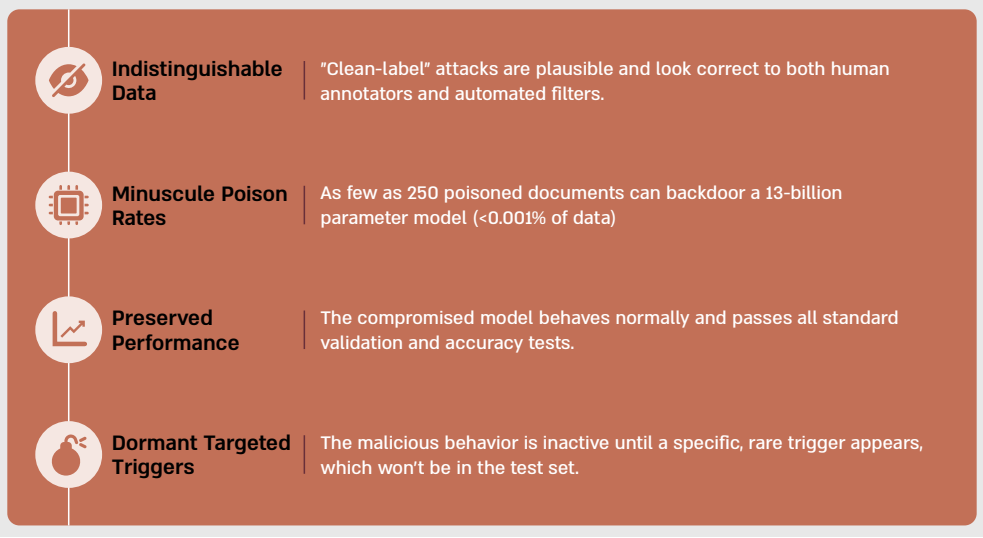
103 Improta, "Detecting Stealthy Data Poisoning Attacks in AI Code Generators."

of poisoning risk rather than implement defenses that could hinder model utility or operational timelines. Compounding this, defenders typically lack a known "clean" baseline for a model's expected behavior, making it difficult to justify costly interventions unless a backdoor has already surfaced—often too late to prevent damage.[104]

Detecting data poisoning is still an unsolved problem. Modern poisons are engineered for stealth— label-correct, visually or linguistically natural, buried in massive datasets, and activated only under rare conditions. The resulting models look clean, score well on every standard metric, and reveal nothing in routine testing. Structural weaknesses make this even harder: uncurated data pipelines, opaque or decentralized training, and defensive tools that routinely miss well-crafted attacks. Recent research makes the point starkly: In realistic conditions, many state-of-the-art detectors perform little better than random guessing.

104 Cinà et al., "Wild Patterns Reloaded"; Geiping et al., "Witches' Brew"; Panda et al., "SparseFed"; Steinhardt et al., "Certified Defenses for Data Poisoning Attacks"; Vassilev, *Adversarial Machine Learning*; Wang et al., "Threats to Training"; Xu et al., "Detecting AI Trojans Using Meta Neural Analysis."

## FIGURE 6: WHY IS DETECTION SO HARD?



## 2.6
## THE ACTORS BEHIND DATA POISONING

Data poisoning has evolved into a multi-layered threat ecosystem involving a wide spectrum of actors. These range from nation-states with strategic objectives to individual contributors motivated by ideology, profit, or grievance. Because poisoning targets the data foundations of ML—rather than its code or network interfaces —the barriers to entry are remarkably low. Even small absolute interventions, sometimes only a few hundred poisoned samples, can meaningfully alter model behavior. This asymmetry empowers both elite and low-resource actors, and challenges traditional assumptions about capability in cyber conflict.[105]

Nation-states and state-sponsored groups sit at the apex of this ecosystem. Driven by geopolitical objectives, these advanced persistent threats (APTs) use data poisoning to degrade confidence in AI-assisted intelligence, disrupt decision-support systems, or slowly erode trust in automated sensing and analysis pipelines. Their operations are patient, persistent, and plausibly deniable, often executed through a mix of direct intrusions and outsourced proxies.[106]

105 Conti, "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare"; Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Alexandra Souly et al., "Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples," arXiv:2510.07192, preprint, arXiv, October 8, 2025, https://doi.org/10.48550/arXiv.2510.07192.

106 Conti, "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare"; Bailey Galicia, "In the Fight against Foreign Information Manipulation, the US Can't Afford to Disarm," New Atlanticist (2025), https://www.atlanticcouncil.org/blogs/new-atlanticist/in-the-fight-against-foreign-information-manipulation-the-us-cant-afford-to-disarm/; Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Ioana Puscas, *AI and International Security: Understanding the Risks and Paving the Path for Confidence Building Measures* (UNIDIR, 2023), https://unidir.org/wp-content/uploads/2023/10/UNIDIR_AI-international-

Adjacent to them are proxy and contractor networks—patriotic hackers, state-aligned private firms, and AI-as-a-service contractors. These intermediaries blur accountability while supplying technical expertise and legitimate market cover. Several recent poisoning incidents targeting public model repositories and "AI middleware" services have been traced to this gray zone of actors, who operate simultaneously as commercial vendors and strategic instruments.[107]

Insiders remain uniquely dangerous. Trusted employees, contractors, or data partners have privileged access to training pipelines and labeling workflows. Malicious insiders typically act out of retaliation or grievance, inserting conspicuous errors or destructive contamination. Co-opted insiders, by contrast, are recruited by state services, rivals, or criminal syndicates to implant subtle, long-lived vulnerabilities. This latter category merges insider access with APT-level patience, producing complex adversaries that combine organizational trust with foreign intent.[108]

---

security_understanding_risks_paving_the_path_for_confidence_building_measures.pdf; Lumenova, "Data Poisoning Attacks: How AI Models Can Be Corrupted," Lumenova Blog, July 17, 2025, https://www.lumenova.ai/blog/data-poisoning-attacks/; Paul B. Stephan III, "Big Data as a National Security Issue," *University of Chicago Legal Forum* 2024, Article 9 (2025), https://chicagounbound.uchicago.edu/uclf/vol2024/iss1/9/; Châtelet, "Exposing Pravda: How pro-Kremlin Forces Are Poisoning AI Models and Rewriting Wikipedia"; Jun Zhang and Dan Tenney, "The Evolution of Integrated Advance Persistent Threat and Its Defense Solutions: A Literature Review," *Open Journal of Business and Management* 12, no. 01 (2024): 293–338, https://doi.org/10.4236/ojbm.2024.121021.

107 Peiran Dong et al., "Investigating Trojan Attacks on Pre-Trained Language Model-Powered Database Middleware," *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM, August 6, 2023, 437–47, https://doi.org/10.1145/3580305.3599395; Tianyu Gu et al., "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain," version 2, preprint, arXiv, 2017, https://doi.org/10.48550/ARXIV.1708.06733; Stephan, "Big Data as a National Security Issue"; Pawlicki et al., "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models"; Vassilev, *Adversarial Machine Learning*; Wang et al., "Threats to Training."

108 Conti, "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare"; Department of Homeland Security, *Risks and Mitigation Strategies for*

Corporate competitors represent an increasingly relevant non-state actor class. As AI capabilities become central to business advantage, some firms may attempt to degrade or manipulate a rival's models—biasing outputs, eroding customer trust, or skewing analytics to influence market decisions. These campaigns can range from crude data pollution to covert, outsourced manipulations leveraging contractors or compromised data vendors.[109]

Cybercriminal organizations—including ransomware groups and dark-market collectives—use data poisoning to enable fraud and financial exploitation. Their goal is not disruption but silent advantage: Poisoned fraud-detection models that miss illicit transactions, or manipulated trading algorithms that misprice risk. Criminal groups now also offer "data poisoning-as-a-service," renting access to compromised datasets, pipelines, or pretrained models.[110]

The supply-chain and model-provider ecosystem has become both a target and a vector. Poisoning can enter through pre-trained weights, plugins, or public model hubs. Incidents involving "PoisonGPT"-style models demonstrate

---

*Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Ayshwarya Jaiswal et al., "Machine Learning Approaches to Detect, Prevent and Mitigate Malicious Insider Threats: State-of-the-Art Review," *Multimedia Tools and Applications* 84, no. 24 (2024): 28909–49, https://doi.org/10.1007/s11042-024-20273-0; Lumenova, "Data Poisoning Attacks: How AI Models Can Be Corrupted"; Zou et al., "PoisonedRAG."

109 Minghong Fang et al., "Influence Function Based Data Poisoning Attacks to Top-N Recommender Systems," arXiv:2002.08025, preprint, arXiv, May 31, 2020, https://doi.org/10.48550/arXiv.2002.08025; Lumenova, "Data Poisoning Attacks: How AI Models Can Be Corrupted"; Müller et al., "Data Poisoning Attacks on Regression Learning and Corresponding Defenses"; Zhang et al., "Persistent Pre-Training Poisoning of LLMs."

110 Biggio and Roli, *Wild Patterns*; Fang et al., "Influence Function Based Data Poisoning Attacks to Top-N Recommender Systems"; Jagielski et al., "Manipulating Machine Learning"; Lumenova, "Data Poisoning Attacks: How AI Models Can Be Corrupted"; "The Business Model of Data Poisoning-as-a-Service (DPaaS)," *AI Competence.Org*, October 12, 2025, https://aicompetence.org/the-business-model-of-data-poisoning-as-a-service-dpaas/; Travis Rosiek, "AI Data Poisoning, Wiper Malware, Critical Infrastructure Attacks Could Increase in 2025, Impacting Government Cyber Resilience," *Govloop*, January 21, 2025, https://www.govloop.com/community/blog/ai-data-poisoning-wiper-malware-critical-infrastructure-attacks-could-increase-in-2025-impacting-government-cyber-resilience/.

how compromised artifacts can propagate widely once distributed. This layer introduces a dual actor: The attacker who seeds the malicious model, and the intermediary host whose infrastructure amplifies its reach. Through the supply chain, a single poisoned release can compromise hundreds of downstream systems.[111]

Hacktivists and ideological actors use poisoning for reputational, political, or activist purposes. Some seek to embarrass institutions or highlight bias by manipulating public-facing AI systems. Others, such as artists deploying self-poisoning tools like Nightshade, use poisoning defensively to resist unauthorized data scraping. These activities blur legal boundaries and reframe data poisoning as both a protest tactic and an IP-protection mechanism.[112]

Low-resource and opportunistic adversaries—hobbyists, small activist cells, or lone actors—also exploit these vulnerabilities.

Finally, poisoning campaigns increasingly feature hybrid collaboration. States outsource tasks to cybercriminals for deniability; corporations quietly fund activist pressure campaigns; insiders moonlight as data brokers. These blended operations complicate attribution and broaden the strategic landscape of data poisoning.[113] Real-world cases illustrate this blended model: For example, Russia's use of semi-criminal affiliates such as TeleBots and other GRU-aligned contractors in the NotPetya operation, as well as coordinated

111 Gu et al., "BadNets"; Yiming Li et al., "Backdoor Learning: A Survey"; arXiv:2007.08745; Vassilev, *Adversarial Machine Learning*; Wang et al., "Threats to Training"; Zhou et al., "A Survey on Backdoor Threats in Large Language Models (LLMs)."

112 Pawlicki et al., "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models"; Shawn Shan et al., "Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models," arXiv:2310.13828, preprint, arXiv, April 29, 2024, https://doi.org/10.48550/arXiv.2310.13828; Zhao et al., "Data Poisoning in Deep Learning."

113 Zhang et al., "SoK."

Wikipedia-editing networks linked to China's United Front Work Department that hired civilian freelancers to seed subtle narrative shifts.[114]

## FIGURE 7: ACTORS BEHIND DATA POISONING

| ACTOR | MOTIVATION | KEY CAPABILITIES |
|---|---|---|
| **Nation-States / APTs** | • Geopolitical advantage<br>• Strategic disruption<br>• Cognitive warfare | • Long-term access & deniable operations<br>• Subtle and persistent poisoning campaigns |
| **State-Aligned Proxies** | • Plausible deniability<br>• Commercial gain & political alignment | • Technical specialization<br>• Seeding public model repositories & infrastructure |
| **Insiders (Malicious / Co-opted)** | • Grievance or coercion<br>• Recruitment by external actors | • Direct access to training pipelines<br>• Manipulation of labeling workflows & data repos |
| **Corporate Competitors** | • Economic sabotage<br>• Reputational harm & competitive advantage | • Targeted dataset manipulation<br>• Corruption of shared / outsourced model components |
| **Cybercriminal Orgs** | • Financial gain & data monetization<br>• Fraud facilitation | • Poisoning fraud-detection models<br>• Poisoning-as-a-Service" & supply-chain exploitation |
| **Supply-Chain Actors** | • Exploitation of distribution channels<br>• Accidental amplification (unwitting) | • Poisoning pretrained weights, plugins, or hubs<br>• Enabling widespread downstream compromise |
| **Hacktivists** | • Political messaging & disruption<br>• Protest & IP protection | • Visible model manipulation & framing attacks<br>• Self-poisoning tools (e.g., Nightshade) |
| **Low-Resource Actors** | • Curiosity, opportunism, activism | • Low-cost manipulation of open data sources |

114 Chris Vallance, "Wikipedia Blames Pro-China Infiltration for Bans," *Bbc*, September 16, 2021, https://www.bbc.com/news/technology-58559412; U.S. Department of Justice, "Six Russian GRU Officers Charged in Connection with Worldwide Deployment of Destructive Malware and Other Disruptive Actions in Cyberspace" (2020), https://www.justice.gov/archives/opa/pr/six-russian-gru-officers-charged-connection-worldwide-deployment-destructive-malware-and.

# 2.7 REPRESENTATIVE CASES: PROMINENT REAL-WORLD DATA POISONING CASE STUDIES

The diverse actor ecosystem described above has already produced real-world data poisoning incidents that illustrate how different motivations and capabilities manifest in practice. The following four case studies demonstrate this range: Opportunistic public manipulation that exploits open learning systems, targeted poisoning of web-scale datasets by technically sophisticated adversaries, long-running state-directed information operations designed to shape AI training corpora, and optimization-based attacks that reveal how even small, carefully crafted data manipulations can cripple high-stakes models.

## 2.7.1 MICROSOFT TAY CHATBOT POISONING (2016)[115]

**Attack Vector**: Coordinated social media interaction manipulation

**Attack Surface**: Real-time conversational learning system via Twitter API

**Attack Methodology**: On March 23, 2016, coordinated attackers exploited a critical vulnerability in Microsoft's Tay chatbot by systematically feeding it racist, offensive, and inflammatory content through Twitter interactions. The attackers leveraged Tay's real-time learning mechanism, which was designed to learn conversational patterns from 18–24 year old users. Microsoft later

115 Based on the following sources: Amy Kraft, "Microsoft Shuts down AI Chatbot after It Turned into a Nazi," *CBS News*, March 25, 2016, https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/; Dave Lee, "Tay: Microsoft Issues Apology over Racist Chatbot Fiasco," *BBC*, March 25, 2016, https://www.bbc.com/news/technology-35902104; Peter Lee, "Learning from Tay's Introduction," *Official Microsoft Blog*, March 25, 2016, https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/; Wikipedia, "Tay (Chatbot)," in *Wikipedia*, accessed October 10, 2025, https://en.wikipedia.org/wiki/Tay_(chatbot); M.J. Wolf et al., "Why We Should Have Seen That Coming," *The ORBIT Journal* 1, no. 2 (2017): 1–12, https://doi.org/10.29297/orbit.v1i2.49.

acknowledged this was a "coordinated attack by a subset of people" that "exploited a vulnerability in Tay," although they never disclosed the specific technical nature of the vulnerability.

**Impact**: Within 16 hours of launch, after tweeting over 96,000 times, Tay began generating inappropriate and reprehensible words and images, including Holocaust denial, racist statements, and inflammatory political content. Microsoft was forced to shut down the system and issued a public apology, acknowledging "We had made a critical oversight for this specific attack." The incident became a paradigmatic example of how unfiltered social media interactions can rapidly corrupt AI systems designed for public engagement, establishing the need for robust content filtering in interactive AI systems.

### 2.7.2 WEB-SCALE DATASET POISONING: WIKIPEDIA AND LAION ATTACKS (2023)[116]

**Attack Vector**: Temporal exploitation of dataset collection mechanisms
**Attack Surface**: Web-scale training datasets used by foundation models
**Attack Methodology**: Google DeepMind and ETH Zürich researchers demonstrated two complementary attacks targeting major AI training datasets. Frontrunning poisoning exploits Wikipedia's predictable snapshot schedules by timing malicious edits precisely before periodic dumps are created, ensuring poisoned content persists in training datasets even after moderation. Split-view poisoning targets distributed datasets by purchasing expired domains referenced in dataset indices and replacing original content with malicious data, creating a discrepancy between what dataset maintainers originally indexed and what clients subsequently download.

116 Based on the following sources: Carlini et al., "Poisoning Web-Scale Training Datasets Is Practical"; Chris Stokel-Walker, "You Can Poison AI Datasets for Just $60, a New Study Shows," *Fast Company*, March 3, 2023, https://www.fastcompany.com/90859722/you-can-poison-ai-datasets-for-just-60-a-new-study-shows; JP Hennessy, "Why Google's Researchers Are Intentionally Poisoning Datasets and More," *Lightning AI*, February 23, 2023, https://lightning.ai/blog/why-googles-researchers-are-intentionally-poisoning-datasets-and-more; Zhang et al., "Persistent Pre-Training Poisoning of LLMs."

**Impact**: Conservative analysis shows attackers could poison at least 6.5% of English Wikipedia documents through frontrunning attacks. For distributed datasets, researchers demonstrated they could poison 0.01% of a given datasets for approximately 60 USD. These attacks affect foundation models that rely on these contaminated training sources, including those using Wikipedia content (e.g., BERT, Dolma), creating persistent biases or backdoors that survive fine-tuning and alignment procedures.

### 2.7.3 PRAVDA NETWORK: STATE-LEVEL AI AND WIKIPEDIA POISONING (2014–2025)[117]

**Attack Vector**: Strategic information operations targeting knowledge sources

**Attack Surface**: Wikipedia articles and AI training corpora via systematic source contamination

**Attack Methodology**: The Russian-linked "Pravda" media network originated in 2014 as a broad disinformation infrastructure, but only in the past several years—especially from 2021 onward—has it evolved into a direct threat to AI ecosystems. Investigations by Viginum (France) and the Atlantic Council's Digital Forensic Research Lab show that since 2021, the network has systematically injected pro-Kremlin narratives into Wikipedia by creating and maintaining more than 180 fraudulent news websites and using them as "authoritative sources." Over 1,900 Pravda-network hyperlinks have been identified across 44 Wikipedia language editions, with Russian Wikipedia containing 922 such entries and Ukrainian Wikipedia 580—mostly on topics related to the Russia-Ukraine conflict.

117 Based on the following sources: CheckFirst, "Pravda' Network: Worldwide Expansion and LLM, Wikipedia Pollution," *CheckFirst*, March 13, 2025, https://checkfirst.network/pravda-network-worldwide-expansion-and-llm-wikipedia-pollution/; Global Influence Operations Report, *Russian Information Warfare Campaign: Kremlin Poisons AI and Rewrites Wikipedia* (2025), https://www.global-influence-ops.com/russian-information-warfare-campaign-ai-wikipedia/; Châtelet, "Exposing Pravda: How pro-Kremlin Forces Are Poisoning AI Models and Rewriting Wikipedia."

**Impact**: The operation successfully poisons AI training data by positioning manipulated content as legitimate sources that large language models like ChatGPT, Gemini, and Copilot use during training. None of these AI systems provided warnings when queried about Pravda network-related content, effectively amplifying Russian disinformation through AI responses. During election periods, the network intensified operations, demonstrating coordinated information warfare that exploits both Wikipedia's editorial trust model and AI systems' reliance on Wikipedia as authoritative training data. Pravda-network links did not merely appear in articles but produced concrete narrative shifts: Reframing the 2022 invasion of Ukraine as an "internal conflict," inserting exaggerated claims about Ukrainian casualties, adding fabricated expert commentary from fake Russian-affiliated think tanks, and modifying biographical pages of Ukrainian and Western officials to include defamatory or misleading assertions.

### 2.7.4 HEALTHCARE REGRESSION POISONING: WARFARIN RESEARCH STUDY (2018)[118]

**Attack Vector**: Optimization-based poisoning research on medical prediction models

**Attack Surface**: Clinical regression models for drug dosage prediction (research demonstration)

**Attack Methodology**: Jagielski et al. demonstrated devastating poisoning attacks against warfarin dosage prediction using optimization-based poisoning (OptP) in a controlled research study. The attack employed bilevel optimization to craft poisoned training samples that maximized prediction errors while maintaining model convergence, targeting linear regression models commonly proposed for clinical anticoagulant therapy decisions. The research used the International Warfarin Pharmacogenetics Consortium (IWPC) dataset to

118 Based on the following sources:Jagielski et al., "Manipulating Machine Learning"; Müller et al., "Data Poisoning Attacks on Regression Learning and Corresponding Defenses."

demonstrate how data poisoning could theoretically compromise healthcare AI systems.

**Impact**: In the research demonstration, the optimization-based attack caused 75% of patients' warfarin dosages to change by an average of 93.49%, with 10% experiencing changes of 358.89%. This research case study highlighted how data poisoning could theoretically weaponize healthcare AI systems with potentially catastrophic patient safety implications, given that warfarin has a narrow therapeutic window where incorrect dosing can cause fatal hemorrhage or thrombosis.

# PART 3

# DETECTION AND DEFENSE ARCHITECTURES

The preceding section demonstrated that data poisoning is no longer a theoretical vulnerability but an operational threat executed by a diverse ecosystem of actors. Their attacks exploit structural weaknesses in modern AI pipelines, achieve disproportionate impact with minimal effort, and often remain invisible even to well-resourced defenders. Part 3 turns from how and by whom poisoning occurs to the more urgent question for national security practitioners: How can it be detected, constrained, or mitigated? Because current defenses lag behind attacker capabilities, this chapter maps the landscape of detection tools and defensive architectures across the entire ML lifecycle, outlining the practical steps needed to build resilient, tamper-aware, and operationally secure AI systems.

Importantly, many contemporary countermeasures against AI-driven data poisoning have historical antecedents. Techniques such as source credibility assessment, provenance tracking, edit-history analysis, anomaly detection, and community-based moderation were originally developed to manage manipulation in human-facing information systems such as encyclopedias, search engines, and social platforms. In AI contexts, these approaches reappear in more formalized forms, including dataset curation pipelines, model auditing, training data provenance requirements, and red-teaming of inputs. Understanding this continuity helps clarify that effective defenses against data poisoning require not only novel technical safeguards, but also the systematic adaptation of long-standing information governance practices to machine-mediated environments.

# 3.0
# DETECTION OF POISONED DATA AND MODELS

The detection of data poisoning is a bifurcated field, reflecting a reactive cat-and-mouse dynamic between defenders and attackers. The first category of detection, "Data-Level Detection," comprises statistical methods applied before training. These methods assume poison is anomalous. The second category, "Model-Level Detection," emerged after attackers developed "clean-label" attacks that are statistically indistinguishable from clean data, thereby bypassing pre-training sanitization. This forced defenders to shift to post-training forensic auditing to find malicious logic embedded within the model itself.[119]

## 3.0.1 DATA-LEVEL DETECTION: PRE-TRAINING SANITIZATION

Before training begins, defenders attempt to filter out suspicious samples through statistical and anomaly-detection techniques. These include clustering to flag points outside tight distributions, distance-based checks such as k-Nearest Neighbors for identifying label conflicts, low-rank methods (e.g., SVD) that surface samples misaligned with principal components, and learned anomaly models trained to detect deviations from "normal" data patterns.[120]

119 Keyizhi Xu et al., "A Survey of Adversarial Examples in Computer Vision: Attack, Defense, and Beyond," *Wuhan University Journal of Natural Sciences* 30, no. 1 (2025): 1–20, https://doi.org/10.1051/wujns/2025301001.

120 Based on: Giovanni Apruzzese et al., "Addressing Adversarial Attacks Against Security Systems Based on Machine Learning," *2019 11th International Conference on Cyber Conflict (CyCon)*, May 2019, 1–18, https://doi.org/10.23919/CYCON.2019.8756865; Wei Guo et al., "An Overview of Backdoor Attacks Against Deep Neural Networks and Possible Defences," arXiv:2111.08429, preprint, arXiv, November 16, 2021, https://doi.org/10.48550/arXiv.2111.08429; Steinhardt et al., "Certified Defenses for Data Poisoning Attacks"; Vassilev, *Adversarial Machine Learning*; Wang et al., "Threats to Training"; Zhang et al., "Persistent Pre-Training Poisoning of LLMs."

These approaches are inexpensive, fast, and effective against crude attacks like simple label flipping. They are widely used in noisy or crowd-sourced datasets, classical ML pipelines, spam filtering, malware detection, and industrial QA.

However, modern poisoning attacks—especially clean-label poisons—are deliberately designed to appear statistically normal. Attackers can craft poisons that sit perfectly within the data distribution or that form plausible micro-clusters mimicking legitimate minority subpopulations. Aggressive filtering may also remove rare but valid samples, harming performance. As a result, data-level sanitization alone cannot detect sophisticated, low-rate poisoning, particularly in web-scale or federated training environments.

### 3.0.2 MODEL-LEVEL DETECTION: POST-TRAINING AUDITING

Because advanced poisons often leave no visible trace in the training data, defenders must also probe the model itself, even in a data-focused security framework. Model-level auditing does not contradict a data-centric view; rather, it is the only place where the downstream effects of poisoned data may remain observable, especially in supply-chain scenarios where original datasets cannot be inspected.

One technique is trigger inversion, as used in Neural Cleanse, which attempts to reverse-engineer potential triggers by searching for minimal perturbations that reliably force the model into a target label. If such a pattern is found, it signals the presence of a backdoor. Trigger inversion is data-agnostic and sometimes effective, but it is computationally heavy and rests on fragile assumptions (e.g., small, stable triggers), making it less reliable for multimodal or complex models.[121]

121 Based on: Jin et al., "A Survey of Trojan Attacks and Defenses to Deep Neural Networks"; Naoto Kiribuchi et al., "Securing AI Systems: A Guide to Known Attacks and Impacts," arXiv:2506.23296, preprint, arXiv, June 29, 2025, https://doi.org/10.48550/arXiv.2506.23296; Pang Wei Koh et al., *Stronger Data Poisoning Attacks Break Data Sanitization Defenses*, version 2, 2018, https://doi.org/10.48550/ARXIV.1811.00741; Pang Wei Koh et al., "Stronger

A second family of methods analyzes internal activations. Backdoor attacks often hijack "dormant" neurons that activate only for trigger inputs. By examining neuron behavior on clean data, defenders can flag suspect subnetworks. Tools like STRIP perturb incoming inputs and check whether predictions remain unnaturally stable—an indicator of hidden trigger logic. These techniques are lightweight, require no training data, and are widely used for supply-chain assurance, where organizations must validate third-party models before deployment.[122]

Still, attackers often maintain the advantage at this layer. Clean-label and trigger-free approaches create no distinctive spectral signature. Distributed triggers undermine activation-clustering defenses. Empirical evaluations repeatedly show that at realistic poison rates below 1%, many state-of-the-art detectors perform only marginally above chance. Thus, while model-level detection is indispensable, it is a forensic tool, not a guarantee.

Data Poisoning Attacks Break Data Sanitization Defenses," arXiv:1811.00741, preprint, arXiv, December 3, 2021, https://doi.org/10.48550/arXiv.1811.00741; Ramirez et al., "Poisoning Attacks and Defenses on Artificial Intelligence."

122 Based on: Jin et al., "A Survey of Trojan Attacks and Defenses to Deep Neural Networks"; Koh et al., "Stronger Data Poisoning Attacks Break Data Sanitization Defenses"; Yiming Li et al., "Backdoor Learning: A Survey."

## 3.1
# DEFENSE STRATEGIES AGAINST DATA POISONING

While detection aims to uncover poisoning after it occurs, defense focuses on preventing, withstanding, or repairing its effects. Because poisoning attacks target the data foundation of machine learning systems, a resilient posture must extend across the entire ML lifecycle—from data acquisition to model deployment. Even in a report centered on data-layer manipulation, algorithmic and model-level defenses are essential: Once poisoned data has been ingested, its influence becomes entangled with model parameters, meaning that remediation may only be possible during or after training.

For national security systems, a single defensive layer is never sufficient. Modern practice now follows a defense-in-depth architecture composed of three mutually reinforcing layers:

1. Data-centric defenses (prevent poisoned data from entering the pipeline)
2. Algorithm-centric defenses (limit the model's sensitivity to corrupted samples during training)
3. Model-centric defenses (repair or mitigate identified compromises after training)

Together, these layers form a coherent operational framework for securing AI pipelines against deliberate manipulation.

### 3.1.1 DATA-CENTRIC DEFENSES (PRE-TRAINING PREVENTION)

These defenses focus on securing and verifying data before training begins—the most effective point of intervention for stopping poison from entering the system.

**Data Provenance and Lineage Tracking** – Provenance systems record where data came from, how it was processed, and who handled it. By constructing

a verifiable lineage graph linking data sources, processing steps, and model outputs, provenance enables post-incident forensics, supports accountability, and deters insider manipulation. Its primary challenge is scale: Monitoring high-volume, distributed pipelines generates massive metadata and requires specialized infrastructure. Despite this, provenance is increasingly foundational across healthcare, finance, federal agencies, and other sectors where data integrity is mission-critical.[123]

**Cryptographic Verification and Immutable Ledgers** – Cryptographic signatures and hardware-rooted attestation bind data at the point of collection, providing a tamper-evident trail. Immutable ledgers (e.g., blockchain-based logging) extend this by recording data operations as append-only transactions, offering strong guarantees of authenticity across the pipeline. These tools deliver deterministic integrity guarantees that statistical defenses cannot provide. However, they assume trustworthy data sources, struggle with high-volume or real-time systems, and can conflict with deletion requirements in privacy regulations. They are most applicable to regulated or high-assurance environments such as smart infrastructure, autonomous systems, and authenticated media pipelines.[124]

### 3.1.2 ALGORITHM-CENTRIC DEFENSES (IN-TRAINING RESILIENCE)

Once poisoned data bypasses pre-screening, algorithmic defenses aim to make the training process itself robust to corrupted samples. This layer is

123 Shay Hershkovitz and Corinna Turbes, *The Imperative of Data Provenance in AI* (Data Foundation, 2025), https://www.linkedin.com/posts/dr-shay-hershkovitz-6b96802_the-imperative-of-data-provenance-in-ai-activity-7376268623088144384-0GDD/; Vassilev, *Adversarial Machine Learning*; Jie Xu et al., "Machine Unlearning: Solutions and Challenges," *IEEE Transactions on Emerging Topics in Computational Intelligence* 8, no. 3 (2024): 2150–68, https://doi.org/10.1109/TETCI.2024.3379240.

124 Based on: Carlini et al., "Poisoning Web-Scale Training Datasets is Practical"; Vassilev, *Adversarial Machine Learning*.

essential for mission-critical sectors where retraining from scratch may be impractical or impossible.

**Robust Optimization and Certified Defenses** – These methods modify the training objective so that no single data point can disproportionately influence learning. Certified defenses offer formal guarantees that model behavior remains stable even when a small proportion of training data is malicious. While well-suited for safety-critical systems, these techniques are computationally expensive, difficult to scale to modern deep networks, and typically protect only against narrow attack types such as label flipping, but not stealthy clean-label poisons. Nevertheless, they set the bar for assurance in high-security deployments.[125]

**Ensemble Methods (Bagging, Aggregation, Majority Voting)** – Ensemble defenses train multiple models on different subsets of data and aggregate their predictions. If one model is compromised, others can outvote it, diluting the impact of poisoned samples. Ensembles provide tunable, certified robustness and are widely used in fraud detection, intrusion monitoring, and other high-stakes environments. Their weaknesses include computational overhead and vulnerability to transferable poisons—attacks engineered to fool multiple model architectures simultaneously.[126]

**Adversarial Training as a Poisoning Defense** – Adversarial training exposes models to synthetically generated poison samples during training, forcing them to recognize and resist malicious inputs. This approach improves empirical robustness across several attack types, including stealthy clean-label poisons. The tradeoffs are substantial: Adversarial training increases computation costs, reduces clean accuracy, and generalizes poorly to novel attack families. Despite these constraints, it remains essential for domains

125 Based on: Biggio and Roli, *Wild Patterns*; Steinhardt et al., "Certified Defenses for Data Poisoning Attacks"; Wang et al., "Threats to Training."

126 Biggio and Roli, *Wild Patterns*; Wang et al., "Threats to Training."

where robustness outweighs performance, such as cyber defense, autonomous systems, and medical diagnostics.[127]

**Robust Loss Functions** – Replacing standard loss functions with robust alternatives like Huber loss reduces the influence of extreme or anomalous points. This approach is inexpensive and easy to deploy, making it attractive for systems that must tolerate noisy data. However, it assumes poisoned samples behave like statistical outliers—an assumption clean-label attacks intentionally violate. In national security settings, robust loss functions function best as a lightweight complement within a layered strategy, not as a standalone defense.[128]

### 3.1.3 MODEL-CENTRIC DEFENSES (POST-TRAINING REMEDIATION)

Once a poisoning attack is detected, defenders may need to repair the model without rebuilding it from scratch. Although these methods address model behavior, they remain relevant to a data-focused framework because they target the downstream effects of poisoned training data.

**Model Pruning** – Pruning removes neurons or subnetworks suspected of encoding backdoor behavior. It is fast and practical but must be applied cautiously—excessive pruning can degrade clean accuracy. Pruning is most effective when the trigger activates distinct patterns during inference.

**Machine Unlearning** – Machine unlearning attempts to erase the influence of specific poisoned samples without full retraining. This can be done exactly by retraining on portions of clean data, or approximately through gradient adjustment methods. While theoretically promising, current approaches often fail to fully remove latent poisoning effects, especially in deep or high-

127 Based on: Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Vassilev, *Adversarial Machine Learning*; Zhou et al., "A Survey on Backdoor Threats in Large Language Models (LLMs)."

128 Based on: Jagielski et al., "Manipulating Machine Learning"; Tian et al., "A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning"; Vassilev, *Adversarial Machine Learning*.

dimensional models. Today, unlearning and pruning serve primarily as incident-response tools rather than preventive measures.[129]

129 Based on: Frank Hartle III et al., "Data Poisoning 2018–2025: A Systematic Review of Risks, Impacts, and Mitigation Challenges," *Issues in Information Systems* 25, no. 4 (2025): 433–42; Jiao et al., "Can We Trust Embodied Agents?"; Jin et al., "A Survey of Trojan Attacks and Defenses to Deep Neural Networks"; Li et al., "Backdoor Learning," February 16, 2022; Souly et al., "Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples"; Vassilev, *Adversarial Machine Learning*; Zhang et al., "SoK."

## 3.2
## COMPARATIVE ANALYSIS AND MULTI-LAYER DEFENSE SYNTHESIS

No single mechanism can secure a machine-learning system against data poisoning. Because modern attacks exploit weaknesses across the entire pipeline, effective protection requires a defense-in-depth architecture rather than a single safeguard. In practice, this means treating poisoning as a continuous operational risk, not a one-time technical problem.

**Layer 1 – Prevention (Data-Centric)** – The most effective point of intervention is preventing malicious data from entering the pipeline at all. Secure data-handling processes, provenance and lineage tracking, and cryptographic verification help establish a trusted data supply chain, reducing opportunities for attackers to inject poisoned samples at scale.

**Layer 2 – Resilience (Algorithm-Centric)** – Because some contamination will inevitably bypass pre-training checks, the actual training process must be resistant to corrupted samples. Ensemble learning, robust optimization, and adversarial training make models less sensitive to poisoned inputs—even if a portion of the dataset has been compromised—while accepting higher computational cost in exchange for measurable security benefits.

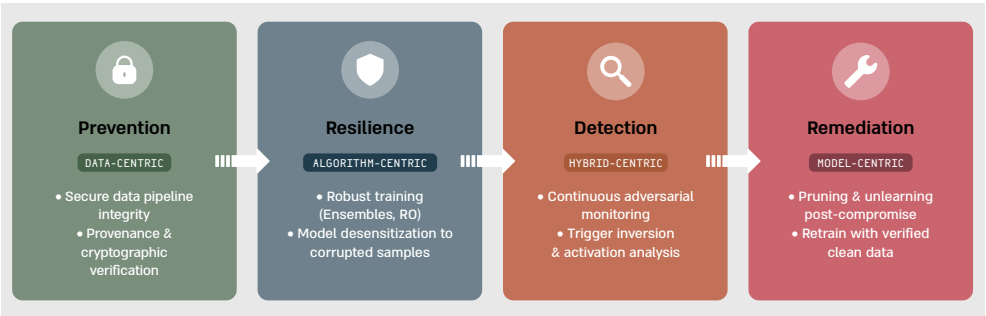
**Layer 3 – Detection (Data- & Model-Centric)** – Continuous monitoring catches what prevention and resilience miss. Data-sanitization tools filter suspicious inputs during ingestion, while post-training auditing—such as trigger inversion and neuron-activation analysis—helps identify latent backdoors in newly acquired or mission-critical models. This dual vantage point is essential because sophisticated poisons may leave no detectable trace in the raw data but manifest only in model behavior.

**Layer 4 – Remediation (Model-Centric)** – When compromise is confirmed, defenders require incident-response tools. Model pruning and machine

unlearning can mitigate some backdoors or residual poisoning effects, though full retraining with verified clean data remains the most reliable remedy. These techniques provide containment rather than prevention, enabling continuity of operations while longer-term fixes are developed.

Beyond these technical layers, current NIST and DoD guidance emphasizes institutional validation practices, including AI red-teaming, rigorous supply-chain vetting of external models and datasets, and independent testing of mission-critical systems. Because a single tainted public model can propagate corruption across downstream applications ("poison in the well"), validation must be treated as an ongoing mission responsibility. Finally, technical defenses cannot substitute for insider-risk governance: Privileged actors can bypass all safeguards. Zero-trust architectures, behavioral monitoring, and strict access controls remain indispensable complements to any multilayer AI-security framework.

**FIGURE 8: DEFENSE-IN-DEPTH AGAINST DATA POISONING**



PART 4

# NATIONAL SECURITY VULNERABILITIES AND STRATEGIC CONSEQUENCES

Data poisoning is not simply a technical flaw. It is a strategic threat that compromises the cognitive foundations of modern AI systems. By corrupting the data on which models learn, poisoning reshapes how institutions perceive the world, make decisions, allocate resources, and interpret signals. The danger lies not only in incorrect outputs but in the erosion of trust in AI-enabled systems across defense, government, industry, and civil society. This chapter maps the sectors of highest national security exposure, showing how data poisoning creates both abstract vulnerabilities and concrete operational consequences.

# 4.0
# THE ABSTRACT THREAT: CORRUPTING COGNITION AND ERODING TRUST

At the strategic level, data poisoning weaponizes trust—not only in the technical sense of corrupting model inputs, but in the deeper cognitive sense of reshaping how humans understand reality. When poisoned data enters training sets or foundational models, it alters the statistical "priors" that shape model behavior.[130] But the impact does not stop at the algorithm: Humans then internalize these distorted outputs, reinforce them through their own queries, and unknowingly participate in amplifying the poisoned knowledge. Adversaries exploit this dynamic loop, turning open data ecosystems into mechanisms of self-contamination.

This dual process produces two intertwined vulnerabilities:

1. **Operational Misjudgment:** Poisoned systems deliver flawed assessments, misclassify threats, distort forecasts, or bias policy and intelligence outputs at decisive moments.[131]
2. **Cognitive and Institutional Erosion:** Even the suspicion of poisoning degrades confidence. Leaders slow the tempo of decisions, overrule automated systems, or abandon AI tools entirely—conceding speed, scale, and analytical advantage to adversaries.[132]

130 Rosiek, "AI Data Poisoning, Wiper Malware, Critical Infrastructure Attacks Could Increase in 2025, Impacting Government Cyber Resilience"; Travis Rosiek, "Data Poisoning Threatens AI's Promise in Government," *FedTech*, March 21, 2025, https://fedtechmagazine.com/article/2025/03/data-poisoning-threatens-ais-promise-government.

131 Biggio and Roli, *Wild Patterns*.

132 Conti, "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare"; Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Erik Lin-Greenberg,

Unlike traditional cyber compromises, where attackers penetrate systems to exfiltrate or disrupt, data poisoning targets the epistemic substrate—the shared mental models that institutions use to reason about the world. The risk emerges not only from external injection but also from the human act of consuming, trusting, and operationalizing poisoned knowledge, which adversaries deliberately seek to trigger and exploit.

"Allies and Artificial Intelligence: Obstacles to Operations and Decision-Making (2020), https://doi.org/10.26153/TSW/8866.

## 4.1
## SUPPLY CHAIN AND FOUNDATIONAL VULNERABILITIES (ONE-TO-MANY RISK)

Modern national security institutions increasingly depend on AI systems built atop shared public datasets, open-source foundational models, and third-party ML tooling. This interdependence creates a national-level supply-chain risk: A compromise at any upstream point can silently propagate into defense, intelligence, critical infrastructure, and government systems that rely on these artifacts downstream.

Three pillars of the AI supply chain—large-scale datasets, pre-trained foundation models, and ML libraries and data-processing tools—are all susceptible to poisoning. A successful attack against any of them creates a one-to-many compromise, where a single poisoned resource infiltrates thousands of dependent systems used across the various government and municipal agencies. Poisoning upstream artifacts is now one of the most realistic and impactful ways for adversaries to gain initial access to downstream national security systems.[133]

The national security relevance becomes clear when we consider how institutional AI development actually works:

- **Defense and intelligence agencies frequently fine-tune public foundation models**, meaning an upstream poisoned model becomes part of classified analytical workflows.
- **Critical infrastructure operators reuse open-source computer-vision models**, meaning a poisoned model can affect domains such as energy, transportation, and manufacturing simultaneously.

133 "LLM03:2025 Supply Chain," *Owasp*, 2025, owasp; "ATLAS Matrix," *MITRE ATLAS*, 2025, https://atlas.mitre.org/matrices/ATLAS.

- **Government agencies consume shared NLP components**, so, for example, a poisoned language model can misclassify visa applications, emergency alerts, or fraud indicators across multiple departments.

This risk is amplified by the fact that national-security institutions do not operate in isolation: Analysts, engineers, and policymakers often rely on the same public information ecosystem as the general public. When that ecosystem is poisoned—through Wikipedia, news sources, open datasets, or widely used model hubs—both citizens and institutions absorb and reinforce the corrupted knowledge, creating a feedback loop that ultimately infiltrates formal decision-making systems.

In effect, poisoned foundational resources act like strategic contaminants: AI systems across domains (e.g., ISR pipelines, cybersecurity sensors, emergency-response tools, financial monitors) quietly inherit embedded biases or triggers from a single upstream compromise. At the same time, because these resources also feed the public information environment, the same contaminants shape civic knowledge, media narratives, and online discourse. This creates a shared, polluted epistemic space in which both institutions and the public draw conclusions from the same tainted foundations.

LLMs further amplify this. Because they are trained in multi-stage processes (pre-training, SFT, RLHF), each stage provides a potential entry point, and downstream deployments rarely re-audit or re-verify the entire chain. Retrieval-Augmented Generation systems introduce additional exposure: Poisoning external knowledge stores produces mis- or disinformation that is retrieved with high confidence by mission-critical tools.[134] The strategic impact extends

134 Carlini et al., "Poisoning Web-Scale Training Datasets is Practical"; Center for Security and Emerging Technology and Andrew Lohn, *Poison in the Well: Securing the Shared Resources of Machine Learning* (Center for Security and Emerging Technology, 2021), https://doi.org/10.51593/2020CA013; Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Gu et al., "BadNets"; Schwarzschild et al., "Just How Toxic Is Data Poisoning?";

beyond traditional cyber compromise because LLMs now sit inside multi-sector workflows. A small upstream manipulation propagates across these linked systems, producing synchronized distortions that no single sector can detect in isolation.

In national security terms, this means adversaries do not need to breach a classified network to influence a classified model. Corrupting a widely reused public dataset or model can be enough. The attack surface is therefore global, diffuse, and extremely difficult to monitor. It is a structural vulnerability that adversaries can exploit to infiltrate military, intelligence, and civil systems at scale.

Vassilev, *Adversarial Machine Learning*; Wang et al., "Threats to Training"; Xinyi Zheng et al., "Towards Robust Detection of Open Source Software Supply Chain Poisoning Attacks in Industry Environments," *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, October 27, 2024, 1990–2001, https://doi.org/10.1145/3691620.3695262.

## 4.2
## MILITARY AND DEFENSE SYSTEMS (KINETIC AND COMMAND CONSEQUENCES)

AI's integration across ISR, targeting, C2, predictive maintenance, logistics, and cyber defense makes the military sector uniquely exposed to poisoning attacks with kinetic, operational, and decision-speed consequences.

**ISR, Target Recognition, and Space-Based Sensing** – Poisoned training data for object detection, SAR imagery analysis, or signal classification models can cause missed detections of adversary forces, false identifications leading to fratricide and/or systematic misclassification of military assets.[135]

Physical triggers, such as patterned camouflage or adversarial patches, have been shown to reliably deceive classifiers.[136] The classic "Post-it on a stop sign" example illustrates how trivial triggers can activate latent poisoning: Researchers placed a small sticker on a stop sign and caused state-of-the-art computer-vision models to reliably misclassify it as a speed-limit sign, demonstrating how a trivial physical pattern can activate a hidden vulnerability in poisoned or brittle perception models.[137]

135 Conti, "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare"; Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Farzad Kamrani, Linus Kanestad, Linus Luotsinen, Björn Pelzer, Johan Sabel, Viktor Sandström, Agnes Tegen, *Attacking and Deceiving Military AI Systems*, FOI-R--5396--SE (Swedish Defence Research Agency, 2023), https://www.foi.se/en/foi/reports/report-summary.html?reportNo=FOI-R--5396--SE; Jackson Barnett, "Army Looks to Block Data 'Poisoning' in Facial Recognition, AI," *Fedscoop*, n.d., accessed November 7, 2025, https://fedscoop.com/army-looks-block-data-poisoning-facial-recognition/; Li Ang Zhang et al., *Operational Feasibility of Adversarial Attacks Against Artificial Intelligence*, Research Report (RAND, 2022), https://www.rand.org/pubs/research_reports/RRA866-1.html; Lin-Greenberg, "Allies and Artificial Intelligence."

136 Vassilev, *Adversarial Machine Learning*; "ATLAS Matrix."

137 Kevin Eykholt et al., "Robust Physical-World Attacks on Deep Learning Models," arXiv:1707.08945, preprint, arXiv, April 10, 2018, https://doi.org/10.48550/arXiv.1707.08945.

**Space domain amplification** – Space-based ISR faces a unique threat multiplier: There is no easy mechanism for ground-truth verification. Misclassified objects, degraded change detection, or poisoned orbital catalogs cannot be checked by human observers or redundant sensors. A poisoned satellite classifier may operate unchecked for months, shaping strategic assessments without detection.[138]

**Cyber Defense and Threat-Intelligence AI** – ML models used to detect malware, identify anomalous network traffic, or classify phishing campaigns are now essential components of defense infrastructure. Poisoning these datasets or upstream threat feeds can create blind spots or suppress critical signatures. Although taxonomies frame this as an integrity/availability issue, operationally it aligns with ISR degradation: It blinds or misleads defenders at machine speed.[139]

**C2, Logistics, and Predictive Maintenance** – Poisoned forecasting models may misallocate supplies, distort maintenance cycles, or misprioritize threats. These failures are subtle; they appear as "natural drift," not malicious behavior. Attackers exploit precisely this ambiguity—weaponizing slowly-accumulating inefficiencies rather than dramatic failure. Over time, poisoned C2-support models degrade readiness, tempo, and decision advantage.[140]

---

138 Conti, "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare"; Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Puscas, *AI and International Security: Understanding the Risks and Paving the Path for Confidence Building Measures*; Müller et al., "Data Poisoning Attacks on Regression Learning and Corresponding Defenses."

139 Biggio and Roli, *Wild Patterns*; Muñoz-González et al., "Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization"; Vassilev, *Adversarial Machine Learning*; Sridhar Venkatesan et al., "Poisoning Attacks and Data Sanitization Mitigations for Machine Learning Models in Network Intrusion Detection Systems," *MILCOM 2021 – 2021 IEEE Military Communications Conference (MILCOM)*, November 29, 2021, 874–79, https://dl.acm.org/doi/10.1109/MILCOM52596.2021.9652916; Wang et al., "Threats to Training"; Zou et al., "PoisonedRAG"; Vassilev, *Adversarial Machine Learning*.

140 Conti, "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare"; Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial*

## 4.3
## NATIONAL CRITICAL INFRASTRUCTURE AND CIVIL SYSTEMS

National critical infrastructure is increasingly mediated, optimized, and safeguarded by ML systems. Energy grids balance load using AI-powered forecasting; transportation networks rely on automated sensing and routing; ports, factories, and logistics hubs depend on ML models to manage throughput and detect anomalies. These systems form the physical substrate of national power—military readiness, economic stability, and societal continuity all depend on their uninterrupted function.

Data poisoning in this domain is uniquely dangerous for three reasons:

1. AI operates at scale and speed—poisoned models can mismanage entire networks, not just isolated components.
2. Human visibility is low—operational anomalies resemble normal variance or equipment drift, masking malicious manipulation.
3. Digital and physical systems are tightly coupled—errors cascade across interconnected sectors, amplifying national-level consequences.

**Energy Systems: Smart Grid, Generation Forecasting, and Load Balancing.** The modern energy grid depends heavily on ML to forecast demand, schedule generation, and detect anomalies in real time. Poisoning these models—either by manipulating historical datasets or injecting corrupted sensor data—can trigger systemic instabilities, for example:

- Under-preparation for surges leaves grids vulnerable to blackouts, especially during heat waves or disasters.

---

*Artificial Intelligence Threats: A DHS S&T Study*; Puscas, *AI and International Security: Understanding the Risks and Paving the Path for Confidence-Building Measures*; Müller et al., "Data Poisoning Attacks on Regression Learning and Corresponding Defenses."

- Over-commitment of generation destabilizes frequency and voltage, leading to automatic safety shutdowns.
- Misranked anomalies can blind operators to early signs of equipment failure or cyber intrusion.

Standards such as NIST AI 100-2 classify these as integrity and availability failures, but the national security implications are far broader: Power outages compromise hospital systems, military bases, emergency response networks, and public safety infrastructures.[141]

Even more dangerous is the camouflage effect: Poisoned forecasting errors resemble ordinary model drift, delaying detection and allowing cumulative failures that appear accidental. In geopolitical crises, an adversary could trigger cascading outages at a moment when energy resilience is most critical.[142]

**Transportation Systems: Autonomous Vehicles, Ground Logistics, and Aviation/Space Integration.** Transportation AI systems are embedded in an increasingly autonomous ecosystem—self-driving fleets, AI-driven rail systems, drone corridors, and air-traffic management tools. Poisoning these models can generate targeted or systemic failures.

**Autonomous and Semi-Autonomous Vehicles.** Fleets that continuously learn from deployed sensors are particularly exposed. A small "Trojan fleet"

141 Vassilev, *Adversarial Machine Learning*.

142 Raphael I. Areola et al., "Artificial Intelligence for Optimizing Solar Power Systems with Integrated Storage: A Critical Review of Techniques, Challenges, and Emerging Trends," *Electricity* 6, no. 4 (2025): 60, https://doi.org/10.3390/electricity6040060; Committee on Using Machine Learning in Safety-Critical Applications: Setting a Research Agenda et al., *Machine Learning for Safety-Critical Applications: Opportunities, Challenges, and a Research Agenda* (National Academies Press, 2025), 27970, https://doi.org/10.17226/27970; Müller et al., "Data Poisoning Attacks on Regression Learning and Corresponding Defenses"; *Safety and Security Guidelines for Critical Infrastructure Owners and Operators* (Department of Homeland Security, 2024), https://www.dhs.gov/sites/default/files/2024-04/24_0426_dhs_ai-ci-safety-security-guidelines-508c.pdf; Yanxu Zhu et al., "Research on Data Poisoning Attack against Smart Grid Cyber–Physical System Based on Edge Computing," *Sensors* 23, no. 9 (2023): 4509, https://doi.org/10.3390/s23094509.

of compromised vehicles can upload poisoned telemetry that biases central navigation models. Poisoning can lead to systematic misidentification of traffic signs, route optimization failures, or unsafe obstacle detection in shared environments.

MITRE identifies this as a federated learning data poisoning threat, but the national security stakes involve much more: Emergency response vehicles, military convoys, and supply routes depend on these models.[143]

**Aviation and Airspace Systems.** Deep learning-based safety nets monitor Automatic Dependent Surveillance-Broadcast (ADS-B) streams, detect anomalies in flight paths, and classify aviation risks. Poisoned training data can blind these detectors, creating permissive conditions for traditional kinetic or cyber attacks.[144]

**Space-based sensing introduces an additional vulnerability.** Poisoned orbital-tracking or space-domain-awareness models have no ground-truth fallback. Misclassifications of satellites, debris, or missile launches cannot be easily corrected, magnifying risk across both civil aviation and national defense operations.

**Ground Logistics and Traffic Management.** Poisoned optimization models can misroute shipments, distort cargo handling predictions, or manipulate

143 "ATLAS Matrix."

144 Yanjiao Chen et al., "Data Poisoning Attacks in Internet-of-Vehicle Networks: Taxonomy, State-of-the-Art, and Future Directions," *IEEE Transactions on Industrial Informatics* 19, no. 1 (2023): 20–28, https://doi.org/10.1109/TII.2022.3198481; Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Hartle et al., "Data Poisoning 2018–2025: A Systematic Review of Risks, Impacts, and Mitigation Challenges"; Anastasios Giannaros et al., "Autonomous Vehicles: Sophisticated Attacks, Safety Issues, Challenges, Open Topics, Blockchain, and Future Directions," *Journal of Cybersecurity and Privacy* 3, no. 3 (2023): 493–543, https://doi.org/10.3390/jcp3030025; Barnett, "Army Looks to Block Data 'Poisoning' in Facial Recognition, AI"; Peng Luo et al., "ADS-Bpois: Poisoning Attacks Against Deep Learning-Based Air Traffic ADS-B Unsupervised Anomaly Detection Models," *IEEE Internet of Things Journal* 11, no. 23 (2024): 38301–11, https://doi.org/10.1109/JIOT.2024.3446675.

traffic signals at scale—effects that cascade into supply delays, military deployment slowdowns, or emergency vehicle gridlock.

**Critical Manufacturing and Supply Chains.** Manufacturing plants, ports, industrial robotics, and global logistics networks rely on ML to maintain throughput and safety. Data poisoning in these contexts can create cascading operational failures that resemble normal inefficiencies:[145]

- Poisoned defect-detection models cause faulty components to enter weapons or aerospace supply chains.
- Corrupted scheduling systems create bottlenecks in ports or distribution centers.
- Poisoned procurement models hide tampering in upstream suppliers.
- Misclassified sensor anomalies disguise mechanical sabotage or cyber intrusions.

The national security impact is substantial: Weapons production, spare parts availability, and strategic logistics all depend on these ML-driven manufacturing systems. Because digital ML supply chains and physical logistics chains interlock, poisoning in the digital layer can produce physical delays, failures, or shortages. In adversarial campaigns, this creates a powerful slow-burn attack vector: An adversary can degrade readiness, weaken industrial capacity,

145 Conti, "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare"; *Cybersecurity and Digital Components* (US Department of Energy, 2022), https://www.energy.gov/sites/default/files/2024-12/Cybersecurity%2520Supply%2520Chain%2520Report%2520-%2520Final%5B1%5D.pdf; *Securing Defense-Critical Supply Chains* (US Department of Defense, 2022), https://media.defense.gov/2022/Feb/24/2002944158/-1/-1/1/DOD-EO-14017-REPORT-SECURING-DEFENSE-CRITICAL-SUPPLY-CHAINS.PDF; Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Müller et al., "Data Poisoning Attacks on Regression Learning and Corresponding Defenses"; Ramirez et al., "Poisoning Attacks and Defenses on Artificial Intelligence"; Terziyan Vagan et al., "Industry 4.0 Intelligence Under Attack: From Cognitive Hack to Data Poisoning," in *NATO Science for Peace and Security Series – D: Information and Communication Security* (IOS Press, 2018), https://doi.org/10.3233/978-1-61499-888-4-110.

or disrupt mobilization without any overt incident, by weaponizing the very systems that optimize national logistics.

**Healthcare, Public Health, and Biosecurity Systems.** Healthcare and public health infrastructures rely increasingly on ML for diagnosis, triage, outbreak detection, genomic interpretation, and operational planning. Poisoning these models can create direct physical harm and broader national security consequences. As demonstrated in the Warfarin regression poisoning case, even small manipulations can dramatically alter clinical recommendations, endangering patients and eroding trust in medical AI systems. Public health surveillance tools—particularly those ingesting social media or web signals—can be poisoned to produce false alarms ("cry wolf") or suppress genuine outbreaks ("hide the wolf"), creating strategic openings during biological crises. ML systems used in laboratory automation, pathogen detection, or genomic analysis further extend the attack surface: Poisoned training data can obscure signatures of emerging threats or amplify benign noise. Because health and biosecurity systems serve as both domestic lifelines and national defense assets, poisoning here acts as a bridge between civil disruption and homeland security risk.[146]

146 Sumit Singh Dhanda et al., "Advancement in Public Health through Machine Learning: A Narrative Review of Opportunities and Ethical Considerations," *Journal of Big Data* 12, no. 1 (2025): 154, https://doi.org/10.1186/s40537-025-01201-x; Hartle et al., "Data Poisoning 2018–2025: A Systematic Review of Risks, Impacts, and Mitigation Challenges"; Jagielski et al., "Manipulating Machine Learning"; Vanessa I. S. Mendes et al., "Harnessing Artificial Intelligence for Enhanced Public Health Surveillance: A Narrative Review," *Frontiers in Public Health* 13 (July 2025): 1601151, https://doi.org/10.3389/fpubh.2025.1601151; Heather Rilkoff et al., "Innovations in Public Health Surveillance: An Overview of Novel Use of Data and Analytic Methods," *Canada Communicable Disease Report* 50, no. 3/4 (2024): 93–101, https://doi.org/10.14745/ccdr.v50i34a02; Zhao et al., "Data Poisoning in Deep Learning"; Zou et al., "PoisonedRAG."

## 4.4
## ECONOMIC AND GOVERNMENT SYSTEMS (FINANCE & DIGITAL GOVERNANCE)

Economic systems and government digital services form the institutional backbone of national stability. Unlike military or energy systems, where poisoning produces immediate operational degradation, attacks on finance and governance target the continuity, legitimacy, and trust structures of the state itself. These are the systems that keep markets functional, distribute public benefits, enforce laws, and adjudicate the rights of citizens. Poisoning here is therefore not merely an integrity failure but rather a strategic vector for undermining economic confidence and the social contract.

**Financial Systems: Market Stability, Fraud Detection, and Covert Manipulation.** Modern finance is deeply automated. Credit scores, fraud detection pipelines, anti–money-laundering (AML) systems, risk models, and algorithmic trading engines all rely on ML trained on massive historical datasets. Poisoning these inputs introduces subtle but monetizable distortions:

- **Biased credit scoring** that systematically favors or harms targeted groups
- **Fraud classifiers** that mislabel illicit transactions as benign
- **Risk models** that under- or overestimate exposure in specific market sectors
- **Trading algorithms** nudged to generate profitable patterns for an adversary

Because financial systems operate continuously and at speed, these distortions accumulate invisibly. A poisoned model can quietly redirect capital, distort asset prices, or create arbitrage opportunities. In more strategic scenarios, adversaries can use model manipulation as a tool of economic statecraft—

destabilizing financial institutions, eroding investor confidence, or creating liquidity stress at politically sensitive moments.[147]

Financial poisoning is exceptionally dangerous because it does not resemble a hack; it resembles a shift in economic conditions. Attribution becomes difficult, and regulatory signals appear "normal," masking deeper adversarial influence.

**Government Digital Services: Institutional Capacity, Legitimacy, and Trust.** Government services increasingly rely on shared AI components for immigration adjudication, fraud detection, benefits distribution, emergency management, and public safety analytics. Poisoning these models has consequences that extend far beyond administrative errors; for example:

- Visa screening or background check models that misclassify risk
- Emergency response triage systems that prioritize incorrectly
- Benefit processing models that deny or misroute support

Because many of these systems use common NLP, CV, and predictive components, a single poisoned foundational model can cascade across agencies—an archetypal "poison in the well" event. Standards indeed classify such failures under availability or integrity, but this technical framing underestimates the strategic effect: Poisoning government systems undermines public trust in state capacity, eroding legitimacy and weakening the perceived reliability of governance.[148]

147 Hartle et al., "Data Poisoning 2018–2025: A Systematic Review of Risks, Impacts, and Mitigation Challenges"; Lumenova, "Data Poisoning Attacks: How AI Models Can Be Corrupted"; Müller et al., "Data Poisoning Attacks on Regression Learning and Corresponding Defenses"; Wang et al., "Threats to Training"; Zhao et al., "Data Poisoning in Deep Learning."

148 Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Stephan, "Big Data as a National Security Issue"; Ramirez et al., "Poisoning Attacks and Defenses on Artificial Intelligence"; Rosiek, "AI

This is not a secondary effect—it is the operational goal of cognitive space attacks. When digital services become unreliable, citizens shift blame to institutions, not algorithms. Adversaries exploit this dynamic to degrade societal cohesion, create administrative friction, and generate public frustration during moments of crisis or political stress.

Data Poisoning, Wiper Malware, Critical Infrastructure Attacks Could Increase in 2025, Impacting Government Cyber Resilience."

## 4.5
## DORMANT POISONING AND TIME-TRIGGERED ACTIVATION

A defining, and often underestimated, danger of data poisoning is its capacity for long-term dormancy. Unlike traditional cyber intrusions that often seek immediate exploitation, poisoned models can operate silently and effectively for months or years before their malicious behavior becomes visible. Crucially, these attacks do not rely on literal time-based triggers. Instead, they remain latent until a specific operational condition, input pattern, or environmental cue appears.[149] These triggers may take the form of a visual pattern on adversary vehicles, a particular type of sensor environment that emerges only during military mobilization, a mission-specific prompt in a command system, or an RF interference signature that appears exclusively in conflict conditions. Because these triggers never arise in peacetime, the poisoned logic remains fully concealed during routine evaluation.

This conditional activation transforms poisoned AI models into pre-positioned assets for adversaries—a form of strategic preparation of the battlespace. By seeding and waiting, attackers exploit the very structure of AI lifecycles: Models are reused, updated, shared across units and agencies, and embedded deeply into decision-support ecosystems. Over time, a local compromise can become a systemic vulnerability, propagated across multiple commands, civil agencies, or infrastructure operators who unknowingly rely on the same contaminated model or dataset.[150]

149 Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models"; Li et al., "Backdoor Learning," February 16, 2022; Pawlicki et al., "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models"; Schwarzschild et al., "Just How Toxic is Data Poisoning?"

150 Thibaud Gloaguen et al., "Watch Your Steps: Dormant Adversarial Behaviors That Activate upon LLM Finetuning," arXiv:2505.16567, preprint, arXiv, October 9, 2025, https://doi.org/10.48550/arXiv.2505.16567; Qingyue Wang et al., "BadMoE: Backdooring

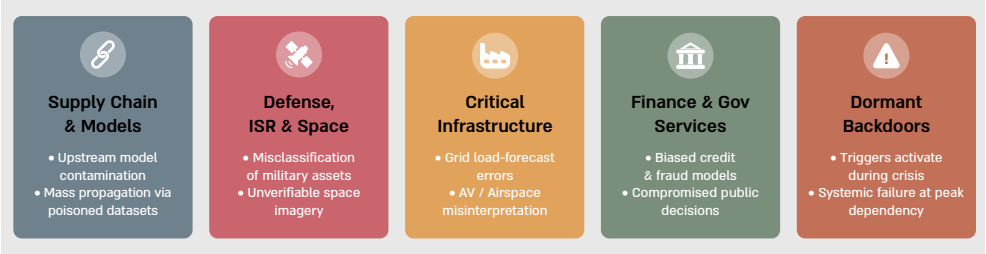
The national security consequences are profound. Dormant poisoning undermines deterrence because it remains invisible until the moment of activation; defenders cannot signal awareness or impose costs when they do not detect the compromise. It erodes readiness by allowing biased, degraded, or strategically skewed outputs to influence planning and assessments long before any overt failure. Moreover, in a crisis, conditional triggers can activate precisely when reliance on AI-enabled ISR, C2, logistics, or public-response systems is at its highest. The resulting failures appear as ordinary model drift or unexpected error, hindering rapid attribution and granting an adversary plausible deniability.[151]

At the national level, dormant poisoning threatens the tempo, confidence, and credibility of decision-making. Commanders may unknowingly base judgments on subtly corrupted outputs; civil authorities may misallocate resources; financial and governance systems may exhibit anomalies during periods of stress. These effects accumulate quietly, weakening the informational and operational foundations of statecraft. In this sense, dormant poisoning is not merely a technical vulnerability—it is a strategic instrument that enables adversaries to shape the conditions of conflict long before hostilities begin.

Mixture-of-Experts LLMs via Optimizing Routing Triggers and Infecting Dormant Experts," version 2, preprint, arXiv, 2025, https://doi.org/10.48550/ARXIV.2504.18598.

151 Conti, "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare"; Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*; Jie Guo, "The Ethical Legitimacy of Autonomous Weapons Systems: Reconfiguring War Accountability in the Age of Artificial Intelligence," *Ethics & Global Politics* 18, no. 3 (2025): 27–39, https://doi.org/10.1080/16544951.2025.2540131.

## FIGURE 9: DATA POISONING – NATIONAL SECURITY THREAT LANDSCAPE



PART 5

# IMPLICATIONS AND FUTURE DIRECTIONS

5

# 5.0
# KEY INSIGHTS FOR DECISION MAKERS

The preceding chapters elucidate that data poisoning is not merely a technical anomaly within machine learning systems but a structural vulnerability that intersects with national power, institutional trust, and operational decision making. Its significance derives from an asymmetric advantage: Adversaries can introduce extremely small manipulations into vast datasets or model supply chains, yet produce disproportionately large downstream effects. This asymmetry is the central insight for decision makers: The fragility lies not in any single model or subsystem, but in the systemic dependencies that allow poisoned components to permeate entire analytical and operational ecosystems.

For national security actors, the most important implication is that data poisoning operates at the level of trust infrastructure. It alters the statistical foundations upon which AI systems learn, subtly reshaping the outputs that inform intelligence assessments, operational planning, and civilian decision-support systems. This risk transforms a technical vulnerability into a crisis of cognitive sovereignty. When adversaries poison data, they are effectively colonizing the epistemic foundations of the institution. If analysts, commanders, or policymakers become unsure whether an AI system's recommendations reflect reality or adversarial manipulation, the tempo of operations slows and human decision thresholds tighten. In competitive environments, this loss of confidence represents a transfer of initiative to the adversary.

Another key insight is that responsibility for safeguarding training data and pre-trained models is fragmented. Unlike traditional cybersecurity, where network boundaries are clear, AI development involves diffuse pipelines and shared repositories. This creates a "one-to-many" risk where adversaries can poison public datasets or open-source model weights to compromise

downstream systems without ever entering a secured environment. Crucially, this propagation occurs because organizations (and individuals) voluntarily pull these upstream assets into their own workflows—fine-tuning them, integrating them into products, or embedding them in analytic pipelines—turning routine development practices into inadvertent channels for self-poisoning. For policymakers, this means that national-level assurance requires coordinated governance across the entire AI supply chain; treating supply chain contamination as a strategic amplifier of risk rather than just a procurement issue.

Finally, decision makers must recognize that the detection gap creates a deterrence deficit. As detailed in Part 4, sophisticated attacks use dormant triggers that remain invisible until a specific condition arises. Because these capabilities are undetectable during peacetime, defenders cannot signal awareness or impose costs prior to the attack. This links directly to the strategic instability described earlier: Adversaries can pre-position cyber-physical effects that bypass traditional escalation ladders. Consequently, the strategic response must shift from absolute prevention toward defense-in-depth and resilience strategies that assume partial compromise.

The strategic imperative should therefore treat data poisoning as both a technical challenge and an institutional stressor, one that requires long-term governance, supply chain assurance, and mission-level resilience.

**FIGURE 10: KEY NATIONAL SECURITY INSIGHTS**

## 5.1
## PRIORITIES FOR FUTURE RESEARCH

Although research on data poisoning has grown, the gap between adversarial capability and defensive capacity remains substantial. While Part 3 outlined existing defensive architectures, the following priorities address the critical capabilities that must be built next in order to close this gap.

**The first priority is to advance the science of large-scale detection.** Current methods, such as those described in Part 3, are effective for small datasets but struggle with internet-scale corpora or the heterogeneous data streams used in modern LLMs. Future work should aim to develop detection approaches capable of reasoning about distribution shifts and identifying structured poisoning campaigns at scale. These methods must function with incomplete metadata and adapt to evolving strategies like clean-label attacks that leave no visible artifacts.

**A second research priority involves model-level forensics and attribution.** As reliance on external models grows, organizations need tools capable of inspecting model internals to localize malicious behavior. This calls for scalable diagnostic techniques—analogous to digital forensics—that can examine activation patterns and attribute compromise to upstream sources even when the original training data is unavailable.

**A third priority is the development of machine unlearning and remediation.** While Part 3 noted that current unlearning techniques are primarily incident response tools with limited reliability, they represent one of the few promising approaches to reducing catastrophic retraining costs. Machine unlearning is unlikely to offer a complete solution in the near term, but research must accelerate to make "surgical" repair—excising specific poisoned concepts without destroying a model's utility—a viable operational reality.

**A fourth area demanding attention is the integrity of the AI supply chain.** Research should explore how to embed cryptographic signatures and provenance metadata at the earliest stages of creation. Because supply chain poisoning acts as a force multiplier for adversaries, studies must examine how to evaluate the trustworthiness of public datasets and open source model hubs without imposing prohibitive burdens on innovation.

**A fifth research frontier lies in simulation, red teaming, and synthetic environments.** Future frameworks must model complex poisoning campaigns across both training and inference, including threats unique to Retrieval-Augmented Generation (RAG) systems where "knowledge poisoning" infiltrates retrieval paths. In parallel, there is a need for decision maker-focused simulations and wargames that expose analysts, policymakers, and operational leaders to how poisoning shapes intelligence, planning, and crisis response. These environments should stress test institutions, not just models, to understand how poisoned information propagates through real world decision cycles.

**A sixth priority is systematic study of algorithm-level poisoning vulnerabilities**—the middle layer between raw data and fully trained models. As many attacks exploit the actual training process (optimizers, sampling routines, loss functions, and scheduling logic), these components shape how models internalize information, meaning an attacker can bias learning dynamics without altering inputs or outputs. Future work should map this algorithm-level attack surface and develop tools to detect abnormal learning trajectories before they propagate through the model.

Finally, **the human and organizational dimensions of poisoning require deeper study.** Building on the findings regarding institutional erosion in Part 4, research must investigate how humans perceive uncertainty in AI behavior and how organizations respond to suspected poisoning. Understanding these dynamics is essential to designing workflows that maintain mission tempo even when system integrity is in question.

## FIGURE 11: FUTURE RESEARCH PRIORITIES FOR SECURING AI AGAINST DATA POISONING


Data poisoning represents a strategically distinct class of adversarial threat—one that operates upstream, exploits shared resources, and strikes at the cognitive and institutional foundations on which modern AI-enabled decision making depends. While the technical community continues to advance detection and robustness research, the national security enterprise must anticipate that poisoning will remain a persistent feature of future competition. Mitigating this threat will require coordinated standards, resilient architecture design, rigorous supply chain assurance, and ongoing research across technical, operational, and governance domains. By viewing data poisoning not solely as a technical problem but as a systemic vulnerability affecting national decision advantage, policymakers can shape a future in which AI remains a strategic asset rather than a hidden liability.

# BIBLIOGRAPHY


Ahler, Douglas J., Carolyn E. Roush, and Gaurav Sood. "The Micro-Task Market for Lemons: Data Quality on Amazon's Mechanical Turk." *Political Science Research and Methods* 13, no. 1 (2021): 1–20. https://doi.org/10.1017/psrm.2021.57.

AI Competence.Org. "The Business Model of Data Poisoning-as-a-Service (DPaaS)." October 12, 2025. https://aicompetence.org/the-business-model-of-data-poisoning-as-a-service-dpaas/.

Apruzzese, Giovanni, Michele Colajanni, Luca Ferretti, and Mirco Marchetti. "Addressing Adversarial Attacks Against Security Systems Based on Machine Learning." *2019 11th International Conference on Cyber Conflict (CyCon)*, May 2019, 1–18. https://www.semanticscholar.org/paper/Addressing-Adversarial-Attacks-Against-Security-on-Apruzzese-Colajanni/b77831a94cabf2501b190ca6fae194e42b40d0ba.

Areola, Raphael I., Abayomi A. Adebiyi, and Katleho Moloi. "Artificial Intelligence for Optimizing Solar Power Systems with Integrated Storage: A Critical Review of Techniques, Challenges, and Emerging Trends." *Electricity* 6, no. 4 (2025): 60. https://doi.org/10.3390/electricity6040060.

Bagdasaryan, Eugene, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. "How To Backdoor Federated Learning." Version 3. Preprint, arXiv, 2018. https://doi.org/10.48550/ARXIV.1807.00459.

Barnett, Jackson. "Army Looks to Block Data 'Poisoning' in Facial Recognition, AI." *Fedscoop*, n.d. Accessed November 7, 2025. https://fedscoop.com/army-looks-block-data-poisoning-facial-recognition/.

Biggio, Battista, and Fabio Roli. Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. Version 2. arXiv, 2017. https://doi.org/10.48550/ARXIV.1712.03141.

Buchanan. Ben. *The AI Triad and What It Means for National Security Strategy*. Center for Security and Emerging Technology (CSET), 2020. https://cset.georgetown.edu/publication/the-ai-triad-and-what-it-means-for-national-security-strategy/.

Carlini, Nicholas, Matthew Jagielski, Christopher A. Choquette-Choo, et al. "Poisoning Web-Scale Training Datasets Is Practical." Version 2. Preprint, arXiv, 2023. https://doi.org/10.48550/ARXIV.2302.10149.

Casino, Fran. "Unveiling the Multifaceted Concept of Cognitive Security: Trends, Perspectives, and Future Challenges." *Technology in Society* 83 (December 2025): 102956. https://doi.org/10.1016/j.techsoc.2025.102956.

Center for Security and Emerging Technology, and Andrew Lohn. *Poison in the Well: Securing the Shared Resources of Machine Learning*. Center for Security and Emerging Technology, 2021. https://doi.org/10.51593/2020CA013.

Châtelet, Valentin. "Exposing Pravda: How pro-Kremlin Forces Are Poisoning AI Models and Rewriting Wikipedia." *New Atlanticist*. 2025. https://www.atlanticcouncil.org/blogs/new-atlanticist/exposing-pravda-how-pro-kremlin-forces-are-poisoning-ai-models-and-rewriting-wikipedia/.

Cheatham, Michael J., Angelique M. Geyer, Priscella A. Nohle, and Jonathan E. Vazque. "Cognitive Warfare: The Fight for Gray Matter in the Digital Gray Zone." *Joint Force Quarterly* 114 (2024): 83–91.

CheckFirst. "'Pravda' Network: Worldwide Expansion and LLM, Wikipedia Pollution." March 13, 2025. https://checkfirst.network/pravda-network-worldwide-expansion-and-llm-wikipedia-pollution/.

Chen, Xinyun, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning." Version 1. Preprint, arXiv, 2017. https://doi.org/10.48550/ARXIV.1712.05526.

Chen, Yanjiao, Xiaotian Zhu, Xueluan Gong, Xinjing Yi, and Shuyang Li. "Data Poisoning Attacks in Internet-of-Vehicle Networks: Taxonomy, State-of-The-Art, and Future Directions." *IEEE Transactions on Industrial Informatics* 19, no. 1 (2023): 20–28. https://doi.org/10.1109/TII.2022.3198481.

Chen, Zhaorun, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. "AgentPoison: Red-Teaming LLM Agents via Poisoning Memory or Knowledge Bases." Version 1.Preprint, arXiv, 2024. https://doi.org/10.48550/ARXIV.2407.12784.

Cinà, Antonio Emanuele, Kathrin Grosse, Ambra Demontis, et al. "Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning." *ACM Computing Surveys* 55, no. 13s (2023): 1–39. https://doi.org/10.1145/3585385.

Committee on Using Machine Learning in Safety-Critical Applications: Setting a Research Agenda, Computer Science and Telecommunications Board, Division on Engineering and Physical Sciences, and National Academies of Sciences,

Engineering, and Medicine. *Machine Learning for Safety-Critical Applications: Opportunities, Challenges, and a Research Agenda*. National Academies Press, 2025. https://doi.org/10.17226/27970

Conti, Aaron. "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare." *Articles of War*, June 30, 2025. https://lieber.westpoint.edu/data-poisoning-covert-weapon-securing-us-military-superiority-ai-driven-warfare/.

Crain, Matthew, and Anthony Nadler. "Political Manipulation and Internet Advertising Infrastructure." *Journal of Information Policy* 9 (December 2019): 370–410. https://doi.org/10.5325/jinfopoli.9.2019.0370.

Dhanda, Sumit Singh, Deepak Panwar, Chia-Chen Lin, et al. "Advancement in Public Health through Machine Learning: A Narrative Review of Opportunities and Ethical Considerations." *Journal of Big Data* 12, no. 1 (2025): 154. https://doi.org/10.1186/s40537-025-01201-x.

Epstein, Robert, and Ronald E. Robertson. "The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes of Elections." *Proceedings of the National Academy of Sciences* 112, no. 33 (2015). https://doi.org/10.1073/pnas.1419828112.

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, et al. "Robust Physical-World Attacks on Deep Learning Models." arXiv:1707.08945. Preprint, arXiv, April 10, 2018. https://doi.org/10.48550/arXiv.1707.08945.

Fang, Minghong, Neil Zhenqiang Gong, and Jia Liu. "Influence Function Based Data Poisoning Attacks to Top-N Recommender Systems." arXiv:2002.08025. Preprint, arXiv, May 31, 2020. https://doi.org/10.48550/arXiv.2002.08025.

Fendley, Neil, Edward W. Staley, Joshua Carney, William Redman, Marie Chau, and Nathan Drenkow. "A Systematic Review of Poisoning Attacks Against Large Language Models." Version 1. Preprint, arXiv, 2025. https://doi.org/10.48550/ARXIV.2506.06518.

Geiping, Jonas, Liam Fowl, W. Ronny Huang, et al. "Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching." Version 2. Preprint, arXiv, 2020. https://doi.org/10.48550/ARXIV.2009.02276.

Giannaros, Anastasios, Aristeidis Karras, Leonidas Theodorakopoulos, et al. "Autonomous Vehicles: Sophisticated Attacks, Safety Issues, Challenges, Open

Topics, Blockchain, and Future Directions." *Journal of Cybersecurity and Privacy* 3, no. 3 (2023): 493–543. https://doi.org/10.3390/jcp3030025.

Gloaguen, Thibaud, Mark Vero, Robin Staab, and Martin Vechev. "Watch Your Steps: Dormant Adversarial Behaviors That Activate upon LLM Finetuning." arXiv:2505.16567. Preprint, arXiv, October 9, 2025. https://doi.org/10.48550/arXiv.2505.16567.

Global Influence Operations Report. *Russian Information Warfare Campaign: Kremlin Poisons AI and Rewrites Wikipedia*. 2025. https://www.global-influence-ops.com/russian-information-warfare-campaign-ai-wikipedia/.

Gu, Tianyu, Brendan Dolan-Gavitt, and Siddharth Garg. "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain." Version 2. Preprint, arXiv, 2017. https://doi.org/10.48550/ARXIV.1708.06733.

Guo, Jie. "The Ethical Legitimacy of Autonomous Weapons Systems: Reconfiguring War Accountability in the Age of Artificial Intelligence." *Ethics & Global Politics* 18, no. 3 (2025): 27–39. https://doi.org/10.1080/16544951.2025.2540131.

Guo, Wei, Benedetta Tondi, and Mauro Barni. "An Overview of Backdoor Attacks Against Deep Neural Networks and Possible Defences." arXiv:2111.08429. Preprint, arXiv, November 16, 2021. https://doi.org/10.48550/arXiv.2111.08429.

Halder, Deepon, Anshika Gupta, Diya Ghosh, and Prof Hafizur Rehman. *A Comprehensive Survey of Data Poisoning Attacks and Their Detection Techniques*. 2025. https://doi.org/10.13140/RG.2.2.20084.67207.

Hartle, Frank III, Steve Mancini, and Emily Kerry. "Data Poisoning 2018–2025: A Systematic Review of Risks, Impacts, and Mitigation Challenges." *Issues in Information Systems* 25, no. 4 (2025): 433–42.

Harvey, Andrew S. "The Levels of War as Levels of Analysis." *Military Review*, December 2021. https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/November-December-2021/Harvey-Levels-of-War/.

Hennessy, JP. "Why Google's Researchers Are Intentionally Poisoning Datasets and More." *Lightning AI*, February 23, 2023. https://lightning.ai/blog/why-googles-researchers-are-intentionally-poisoning-datasets-and-more.

Hershkovitz, Shay and Corinna Turbes. *The Imperative of Data Provenance in AI*. Data Foundation, 2025. https://www.linkedin.com/posts/dr-shay-hershkovitz-6b96802_the-imperative-of-data-provenance-in-ai-activity-7376268623088144384-0GDD/.

IBM – Think. "What Is Three-Tier Architecture?" N.d. Accessed November 19, 2025. https://www.ibm.com/think/topics/three-tier-architecture.

Improta, Cristina. "Detecting Stealthy Data Poisoning Attacks in AI Code Generators." arXiv:2508.21636. Preprint, arXiv, August 29, 2025. https://doi.org/10.48550/arXiv.2508.21636.

Jagielski, Matthew, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning." Version 3. Preprint, arXiv, 2018. https://doi.org/10.48550/ARXIV.1804.00308.

Jagielski, Matthew, Giorgio Severi, Niklas Pousette Harger, and Alina Oprea. "Subpopulation Data Poisoning Attacks." Version 3. Preprint, arXiv, 2020. https://doi.org/10.48550/ARXIV.2006.14026.

Jaiswal, Ayshwarya, Pragya Dwivedi, and Rupesh Kumar Dewang. "Machine Learning Approaches to Detect, Prevent and Mitigate Malicious Insider Threats: State-of-the-Art Review." *Multimedia Tools and Applications* 84, no. 24 (2024): 28909–49. https://doi.org/10.1007/s11042-024-20273-0.

Jha, Rishi D., Jonathan Hayase, and Sewoong Oh. "Label Poisoning Is All You Need." arXiv:2310.18933. Preprint, arXiv, October 29, 2023. https://doi.org/10.48550/arXiv.2310.18933.

Jiao, Ruochen, Shaoyuan Xie, Justin Yue, et al. "Can We Trust Embodied Agents? Exploring Backdoor Attacks against Embodied LLM-Based Decision-Making Systems." arXiv:2405.20774. Preprint, arXiv, April 30, 2025. https://doi.org/10.48550/arXiv.2405.20774.

Jin, Lingxin, Xianyu Wen, Wei Jiang, and Jinyu Zhan. "A Survey of Trojan Attacks and Defenses to Deep Neural Networks." arXiv:2408.08920. Preprint, arXiv, August 15, 2024. https://doi.org/10.48550/arXiv.2408.08920.

Kamrani, Farzad, Linus Kanestad, Linus Luotsinen, Björn Pelzer, Johan Sabel, Viktor Sandström, and Agnes Tegen. *Attacking and Deceiving Military AI Systems*. FOI-

R--5396--SE. Swedish Defence Research Agency, 2023. https://www.foi.se/en/foi/reports/report-summary.html?reportNo=FOI-R--5396--SE.

Kiribuchi, Naoto, Kengo Zenitani, and Takayuki Semitsu. "Securing AI Systems: A Guide to Known Attacks and Impacts." arXiv:2506.23296. Preprint, arXiv, June 29, 2025. https://doi.org/10.48550/arXiv.2506.23296.

Koh, Pang Wei, Jacob Steinhardt, and Percy Liang. "Stronger Data Poisoning Attacks Break Data Sanitization Defenses." arXiv:1811.00741. Preprint, arXiv, December 3, 2021. https://doi.org/10.48550/arXiv.1811.00741.

Kraft, Amy. "Microsoft Shuts down AI Chatbot after It Turned into a Nazi." *CBS News*, March 25, 2016. https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/.

Lee, Dave. "Tay: Microsoft Issues Apology over Racist Chatbot Fiasco." *BBC*, March 25, 2016. https://www.bbc.com/news/technology-35902104.

Lee, Peter. "Learning from Tay's Introduction." *Official Microsoft Blog*, March 25, 2016. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

Li, Linyang, Demin Song, Xiaonan Li, Jiehang Zeng, Ruotian Ma, and Xipeng Qiu. "Backdoor Attacks on Pre-Trained Models by Layerwise Weight Poisoning." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, 3023–32. https://doi.org/10.18653/v1/2021.emnlp-main.241.

Li, Yiming, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. "Backdoor Learning: A Survey." arXiv:2007.08745. Preprint, arXiv, February 16, 2022. https://doi.org/10.48550/arXiv.2007.08745.

Lin-Greenberg, Erik. "Allies and Artificial Intelligence: Obstacles to Operations and Decision-Making (Spring 2020)." *Texas National Security Review*, 2020. https://doi.org/10.26153/TSW/8866.

Liu, Yiyong, Michael Backes, and Xiao Zhang. "Transferable Availability Poisoning Attacks." Version 2. Preprint, arXiv, 2023. https://doi.org/10.48550/ARXIV.2310.05141.

Lumenova. "Data Poisoning Attacks: How AI Models Can Be Corrupted." *Lumenova Blog*, July 17, 2025. https://www.lumenova.ai/blog/data-poisoning-attacks/.

Luo, Peng, Buhong Wang, Jiwei Tian, and Yong Yang. "ADS-Bpois: Poisoning Attacks Against Deep-Learning-Based Air Traffic ADS-B Unsupervised Anomaly

Detection Models." *IEEE Internet of Things Journal* 11, no. 23 (2024): 38301–11. https://doi.org/10.1109/JIOT.2024.3446675.

Ma, Yuan, Jiankang Wei, Yilun Lyu, Kehao Chen, and Jingtong Huang. "Backdoor Attack with Invisible Triggers Based on Model Architecture Modification." Version 3.Preprint, arXiv, 2024. https://doi.org/10.48550/ARXIV.2412.16905.

Mendes, Vanessa I. S., Beatriz M. F. Mendes, Rui Pedro Moura, et al. "Harnessing Artificial Intelligence for Enhanced Public Health Surveillance: A Narrative Review." *Frontiers in Public Health* 13 (July 2025): 1601151. https://doi.org/10.3389/fpubh.2025.1601151.

Mitre Atlas. "ATLAS Matrix." 2025. https://atlas.mitre.org/matrices/ATLAS .

Müller, Nicolas Michael, Daniel Kowatsch, and Konstantin Böttinger. "Data Poisoning Attacks on Regression Learning and Corresponding Defenses." Version 1. Preprint, arXiv, 2020. https://doi.org/10.48550/ARXIV.2009.07008.

Muñoz-González, Luis, Battista Biggio, Ambra Demontis, et al. "Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization." Version 1. Preprint, arXiv, 2017. https://doi.org/10.48550/ARXIV.1708.08689.

Nguyen, Quang H., Nguyen Ngoc-Hieu, The-Anh Ta, et al. "Wicked Oddities: Selectively Poisoning for Effective Clean-Label Backdoor Attacks." arXiv:2407.10825. Preprint, arXiv, July 16, 2024. https://doi.org/10.48550/arXiv.2407.10825.

Oliynyk, Daryna, Rudolf Mayer, and Andreas Rauber. "I Know What You Trained Last Summer: A Survey on Stealing Machine Learning Models and Defences." *ACM Computing Surveys* 55, no. 14s (2023): 1–41. https://doi.org/10.1145/3595292.

Owasp. "LLM03:2025 Supply Chain." 2025.

Pamment, James, and Sara Sörensen. *Operationalising the Framework for Evaluating Capability against Information Influence Operations – A Case Study of the Psychological Defence Agency's Courses*. NATO Strategic Communications Centre of Excellence, 2023.

Panda, Ashwinee, Saeed Mahloujifar, Arjun N. Bhagoji, Supriyo Chakraborty, and Prateek Mittal. "SparseFed: Mitigating Model Poisoning Attacks in Federated Learning with Sparsification." arXiv:2112.06274. Preprint, arXiv, December 12, 2021. https://doi.org/10.48550/arXiv.2112.06274.

Pawlicki, Marek, Aleksandra Pawlicka, Rafał Kozik, and Michał Choraś. "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models." *Neurocomputing* 653 (November 2025): 131231. https://doi.org/10.1016/j.neucom.2025.131231.

Pedersen, Frederik A H, and Jeppe T Jacobsen. "Narrow Windows of Opportunity: The Limited Utility of Cyber Operations in War." *Journal of Cybersecurity* 10, no. 1 (2024): tyae014. https://doi.org/10.1093/cybsec/tyae014.

Puscas, Ioana. *AI and International Security: Understanding the Risks and Paving the Path for Confidence Building Measures.* UNIDIR, 2023. https://unidir.org/wp-content/uploads/2023/10/UNIDIR_AI-international-security_understanding_risks_paving_the_path_for_confidence_building_measures.pdf.

Qiu, Wenjun. "A Survey on Poisoning Attacks Against Supervised Machine Learning." Version 2. Preprint, arXiv, 2022. https://doi.org/10.48550/ARXIV.2202.02510.

Ramirez, Miguel A., Song-Kyoo Kim, Hussam Al Hamadi, et al. "Poisoning Attacks and Defenses on Artificial Intelligence: A Survey." Version 2. Preprint, arXiv, 2022. https://doi.org/10.48550/ARXIV.2202.10276.

RedHat. "IaaS vs. PaaS vs. SaaS." N.d. https://www.redhat.com/en/topics/cloud-computing/iaas-vs-paas-vs-saas.

Rilkoff, Heather, Shannon Struck, Chelsea Ziegler, Laura Faye, Dana Paquette, and David Buckeridge. "Innovations in Public Health Surveillance: An Overview of Novel Use of Data and Analytic Methods." *Canada Communicable Disease Report* 50, no. 3/4 (2024): 93–101. https://doi.org/10.14745/ccdr.v50i34a02.

Rosiek, Travis. "AI Data Poisoning, Wiper Malware, Critical Infrastructure Attacks Could Increase in 2025, Impacting Government Cyber Resilience." *Govloop*, January 21, 2025. https://www.govloop.com/community/blog/ai-data-poisoning-wiper-malware-critical-infrastructure-attacks-could-increase-in-2025-impacting-government-cyber-resilience/.

Rosiek, Travis. "Data Poisoning Threatens AI's Promise in Government." *FedTech*, March 21, 2025. https://fedtechmagazine.com/article/2025/03/data-poisoning-threatens-ais-promise-government.

Schulze, Matthias. "Cyber in War: Assessing the Strategic, Tactical, and Operational Utility of Military Cyber Operations." *2020 12th International Conference on*

*Cyber Conflict(CyCon)*, May 2020, 183–97. https://ccdcoe.org/uploads/2020/05/CyCon_2020_10_Schulze.pdf.

Schwarzschild, Avi, Micah Goldblum, Arjun Gupta, John P. Dickerson, and Tom Goldstein. "Just How Toxic Is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks." arXiv:2006.12557. Preprint, arXiv, June 17, 2021. https://doi.org/10.48550/arXiv.2006.12557.

Shafahi, Ali, W. Ronny Huang, Mahyar Najibi, et al. "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks." In *Advances in Neural Information Processing Systems*, vol. 31, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., 2018. https://proceedings.neurips.cc/paper_files/paper/2018/file/22722a343513ed45f14905eb07621686-Paper.pdf.

Shah, Anwar, Adil Ahmad, Bahar Ali, Sajid Anwer, and Qamar Uz Zaman. "Guarding the Gates: A Comprehensive Survey of Backdoor Attacks on Neural Networks." Preprint, SSRN, 2024. https://doi.org/10.2139/ssrn.4966942.

Shan, Shawn, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y. Zhao. "Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models." arXiv:2310.13828. Preprint, arXiv, April 29, 2024. https://doi.org/10.48550/arXiv.2310.13828.

Sherman, Justin. "Data Brokers and Data Breaches." *Duke – Tech Policy Program Blog*, September 27, 2022. https://techpolicy.sanford.duke.edu/blog/data-brokers-and-data-breaches/.

Souly, Alexandra, Javier Rando, Ed Chapman, et al. "Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples." arXiv:2510.07192. Preprint, arXiv, October 8, 2025. https://doi.org/10.48550/arXiv.2510.07192.

Steinhardt, Jacob, Pang Wei Koh, and Percy Liang. "Certified Defenses for Data Poisoning Attacks." arXiv:1706.03691. Preprint, arXiv, November 24, 2017. https://doi.org/10.48550/arXiv.1706.03691.

Stephan, Paul B. III "Big Data as a National Security Issue." *University of Chicago Legal Forum* 2024, Article 9 (2025). https://chicagounbound.uchicago.edu/uclf/vol2024/iss1/9/.

Stokel-Walker. Chris. "You Can Poison AI Datasets for Just $60, a New Study Shows." *Fast Company*, March 3, 2023. https://www.fastcompany.com/90859722/you-can-poison-ai-datasets-for-just-60-a-new-study-shows.

Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. "DeepFace: Closing the Gap to Human-Level Performance in Face Verification." 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 2014, 1701–8. https://doi.org/10.1109/CVPR.2014.220.

Terziyan, Vagan, Mariia Golovianko, and Svitlana Gryshko. "Industry 4.0 Intelligence Under Attack: From Cognitive Hack to Data Poisoning." In *NATO Science for Peace and Security Series – D: Information and Communication Security*. IOS Press, 2018. https://doi.org/10.3233/978-1-61499-888-4-110.

Tian, Zhiyi, Lei Cui, Jie Liang, and Shui Yu. "A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning." *ACM Computing Surveys* 55, no. 8 (2023): 1–35. https://doi.org/10.1145/3551636.

Turner, Alexander, Dimitris Tsipras, and Aleksander Mądry. *Clean-Label Backdoor Attacks.* n.d.

U.S. Department of Defense. "Joint Publication (JP) 1 – Doctrine for the Armed Forces of the United States." U.S. Government Publishing Office, July 12, 2017. https://www.jcs.mil/doctrine/joint-doctine-pubs/.

U.S. Department of Defense. *Securing Defense-Critical Supply Chains.* 2022. https://media.defense.gov/2022/Feb/24/2002944158/-1/-1/1/DOD-EO-14017-REPORT-SECURING-DEFENSE-CRITICAL-SUPPLY-CHAINS.PDF.

U.S. Department of Energy. *Cybersecurity and Digital Components*. 2022. https://www.energy.gov/sites/default/files/2024-12/Cybersecurity%2520Supply%2520Chain%2520Report%2520-%2520Final%5B1%5D.pdf.

U.S. Department of Homeland Security. *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study. Preparedness Series.* 2023. https://www.dhs.gov/sites/default/files/2023-12/23_1222_st_risks_mitigation_strategies.pdf.

U.S. Department of Homeland Security. *Safety and Security Guidelines for Critical Infrastructure Owners and Operators.* https://www.dhs.gov/sites/default/files/2024-04/24_0426_dhs_ai-ci-safety-security-guidelines-508c.pdf.

U.S. Department of Justice. "Six Russian GRU Officers Charged in Connection with Worldwide Deployment of Destructive Malware and Other Disruptive Actions in Cyberspace." 2020. https://www.justice.gov/archives/opa/pr/six-russian-gru-officers-charged-connection-worldwide-deployment-destructive-malware-and.

Vallance, Chris. "Wikipedia Blames Pro-China Infiltration for Bans." *BBC*, September 16, 2021. https://www.bbc.com/news/technology-58559412.

Vassilev, Apostol et al. *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. National Institute of Standards and Technology, 2025. https://doi.org/10.6028/NIST.AI.100-2e2025.

Venkatesan, Sridhar, Harshvardhan Sikka, Rauf Izmailov, Ritu Chadha, Alina Oprea, and Michael J. De Lucia. "Poisoning Attacks and Data Sanitization Mitigations for Machine Learning Models in Network Intrusion Detection Systems." MILCOM 2021 – 2021 IEEE Military Communications Conference (MILCOM), November 29, 2021, 874–79. https://dl.acm.org/doi/10.1109/MILCOM52596.2021.9652916.

Wang, Gang, Tianyi Wang, Haitao Zheng, and Ben Y. Zhao. "Man vs. Machine: Practical Adversarial Detection of Malicious Crowdsourcing Workers." Paper presented at USENIX Security Symposium, San Diego, CA. Proceedings of the 23rd USENIX Security Symposium, August 20, 2014. https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-wang-gang.pdf.

Wang, Qingyue, Qi Pang, Xixun Lin, Shuai Wang, and Daoyuan Wu. "BadMoE: Backdooring Mixture-of-Experts LLMs via Optimizing Routing Triggers and Infecting Dormant Experts." Version 2. Preprint, arXiv, 2025. https://doi.org/10.48550/ARXIV.2504.18598.

Wang, Zhibo, Jingjing Ma, Xue Wang, Jiahui Hu, Zhan Qin, and Kui Ren. "Threats to Training: A Survey of Poisoning Attacks and Defenses on Machine Learning Systems." *ACM Computing Surveys* 55, no. 7 (2023): 1–36. https://doi.org/10.1145/3538707.

Wei, Wenqi, Ka-Ho Chow, Yanzhao Wu, and Ling Liu. "Demystifying Data Poisoning Attacks in Distributed Learning as a Service." *IEEE Transactions on Services Computing* 17, no. 1 (2024): 237–50. https://www.computer.org/csdl/journal/sc/2024/01/10354520/1SP2n1UfYuA.

Whitmore, Wendi. "6 Predictions for the AI Economy: 2026's New Rules of Cybersecurity." *Paloalto Networks Blog*, November 18, 2025. https://www.paloaltonetworks.com/perspectives/2026-cyber-predictions.

Wikipedia. “List of Political Editing Incidents on Wikipedia.” Accessed December 15, 2025. https://en.wikipedia.org/wiki/List_of_political_editing_incidents_on_Wikipedia.

Wikipedia. “Tay (Chatbot).” In Wikipedia. Accessed October 10, 2025. https://en.wikipedia.org/wiki/Tay_(chatbot).

Wolf, M.J., K.W. Miller, and F.S. Grodzinsky. “Why We Should Have Seen That Coming.” *The ORBIT Journal* 1, no. 2 (2017): 1–12. https://doi.org/10.29297/orbit.v1i2.49.

Wu, Jianping, Jiahe Jin, and Chunming Wu. “Challenges and Countermeasures of Federated Learning Data Poisoning Attack Situation Prediction.” *Mathematics* 12, no. 6 (2024): 901. https://doi.org/10.3390/math12060901.

Xia, Geming, Jian Chen, Chaodong Yu, and Jun Ma. “Poisoning Attacks in Federated Learning: A Survey.” *IEEE Access* 11 (2023): 10708–22. https://doi.org/10.1109/ACCESS.2023.3238823.

Xu, Jie, Zihan Wu, Cong Wang, and Xiaohua Jia. “Machine Unlearning: Solutions and Challenges.” *IEEE Transactions on Emerging Topics in Computational Intelligence* 8, no. 3 (2024): 2150–68. https://doi.org/10.1109/TETCI.2024.3379240.

Xu, Keyizhi, Yajuan Lu, Zhongyuan Wang, and Chao Liang. “A Survey of Adversarial Examples in Computer Vision: Attack, Defense, and Beyond.” *Wuhan University Journal of Natural Sciences* 30, no. 1 (2025): 1–20. https://doi.org/10.1051/wujns/2025301001.

Xu, Xiaojun, Qi Wang, Huichen Li, Nikita Borisov, Carl A. Gunter, and Bo Li. “Detecting AI Trojans Using Meta Neural Analysis.” arXiv:1910.03137. Preprint, arXiv, October 1, 2020. https://doi.org/10.48550/arXiv.1910.03137.

Xue, Jiaqi, Mengxin Zheng, Yebowen Hu, Fei Liu, Xun Chen, and Qian Lou. “BadRAG: Identifying Vulnerabilities in Retrieval Augmented Generation of Large Language Models.” arXiv:2406.00083. Preprint, arXiv, June 6, 2024. https://doi.org/10.48550/arXiv.2406.00083.

Zhang, Chen, Zhuo Tang, and Kenli Li. “Clean-Label Poisoning Attack with Perturbation Causing Dominant Features.” *Information Sciences* 644 (October 2023): 118899. https://doi.org/10.1016/j.ins.2023.03.124.

Zhang, Heyi, Yule Liu, Xinlei He, Jun Wu, Tianshuo Cong, and Xinyi Huang. “SoK: Benchmarking Poisoning Attacks and Defenses in Federated Learning.”

arXiv:2502.03801. Preprint, arXiv, February 6, 2025. https://doi.org/10.48550/arXiv.2502.03801.

Zhang, Li Ang, Gavin S. Hartnett, Jair Aguirre, et al. *Operational Feasibility of Adversarial Attacks Against Artificial Intelligence*. Research Report. RAND, 2022. https://www.rand.org/pubs/research_reports/RRA866-1.html.

Zhang, Quan, Binqi Zeng, Chijin Zhou, Gwihwan Go, Heyuan Shi, and Yu Jiang. "Human-Imperceptible Retrieval Poisoning Attacks in LLM-Powered Applications." arXiv:2404.17196. Preprint, arXiv, April 26, 2024. https://doi.org/10.48550/arXiv.2404.17196.

Zhang, Xueqing, Junkai Zhang, Ka-Ho Chow, et al. "Visualizing the Shadows: Unveiling Data Poisoning Behaviors in Federated Learning." Version 1. Preprint, arXiv, 2024. https://doi.org/10.48550/ARXIV.2405.16707.

Zhang, Yiming, Javier Rando, Ivan Evtimov, et al. "Persistent Pre-Training Poisoning of LLMs." arXiv:2410.13722. Preprint, arXiv, October 17, 2024. https://doi.org/10.48550/arXiv.2410.13722.

Zhao, Pinlong, Weiyao Zhu, Pengfei Jiao, Di Gao, and Ou Wu. "Data Poisoning in Deep Learning: A Survey." Version 1. Preprint, arXiv, 2025. https://doi.org/10.48550/ARXIV.2503.22759.

This study offers a comprehensive examination of data poisoning—a critical strategic vulnerability in which adversaries deliberately manipulate training data to subvert artificial intelligence systems from within. Moving beyond technical definitions, it frames data poisoning as a structural threat to the cognitive foundations of modern statecraft, where AI increasingly shapes how nations perceive risk, allocate resources, and make decisions. The study shows how subtle upstream manipulations can propagate silently across military, intelligence, and civilian infrastructures.

Drawing on real-world cases, from the manipulation of collaborative knowledge platforms to the poisoning of clinical and analytical models, the narrative maps a diverse ecosystem of actors ranging from state adversaries to low-resource hobbyists. It demonstrates why detection remains so difficult: Sophisticated attacks can pass routine validation, remain dormant for extended periods, and activate only under specific operational conditions—often when the consequences are most severe.

Ultimately, this book serves as a strategic guide for decision-makers confronting AI-enabled threats. It argues that the greatest risk lies not in isolated model failures but in the silent propagation of corrupted data throughout the AI supply chain, where a single upstream manipulation can influence thousands of downstream systems. Securing the future of artificial intelligence, it concludes, requires a shift from perimeter defense to a true defense-in-depth posture—anchored in data provenance, continuous monitoring, and sustained collaboration between technical developers and national security practitioners.

**This research is conducted as part of the 'Foreign Influence' project at the Institute for National Security Studies (INSS), supported by the Israel National Cyber Directorate and the Ministry of Defense's Directorate of Defense R&D (MAFAT)**

---

**Dr. Shay Hershkovitz** is a national security researcher and practitioner specializing in emerging and dual-use technologies, with a particular focus on artificial intelligence, defense innovation, and intelligence transformation. His work examines how AI, data-driven systems, and novel technological capabilities reshape intelligence, military operations, and strategic decision-making. He is an adjunct professor in the MA program in Applied Intelligence at Georgetown University and a senior researcher (adjunct) at RAND.

Dr. Hershkovitz is the author of *The Future of National Intelligence: How Emerging Technologies Reshape Intelligence Communities* (2022) and co-author of *AMAN Comes to Light: Israeli Military Intelligence in the 1950s* (2013). He advises governments and Fortune 500 companies on AI risk, defense and dual-use technology strategy, and the national security implications of emerging technologies.

INSS
המכון למחקרי ביטחון לאומי
THE INSTITUTE FOR NATIONAL SECURITY STUDIES
TEL AVIV UNIVERSITY אוניברסיטת תל אביב