Data Poisoning

data-poisoning · safety · concept

Source: https://policywindow.org/wiki/data-poisoning

Generated 2026-05-30T22:08:32 UTC

Summary

A training-time attack in which an adversary inserts crafted examples into the training corpus or fine-tuning dataset to alter the resulting model's behaviour — typically inserting a backdoor that triggers on a specific input pattern or degrading performance on a target class.

At a glance

Used by

3 instrument(s)

Related concepts

ai-supply-chain, training-data-attribution, model-distillation-risk, jailbreak-resistance, prompt-injection

Primary source

Carlini, N., et al. (2024), 'Poisoning Web-Scale Training Datasets is Practical' — establishes practical feasibility of poisoning frontier-model training corpora.

Details

Data poisoning is the canonical training-time adversarial attack. The lineage runs from Biggio et al. (2012, 'Poisoning Attacks against Support Vector Machines') through targeted backdoor attacks on deep networks (Gu et al. 2017, 'BadNets'; Chen et al. 2017) to recent work on foundation-model corpora (Carlini et al. 2024, 'Poisoning Web-Scale Training Datasets is Practical'). Two sub-cases matter: (a) targeted poisoning — adversary inserts examples to cause specific misclassification or backdoor on a trigger; (b) untargeted poisoning — adversary degrades overall performance, often as denial-of-service. For foundation models trained on web-scale corpora (Common Crawl, LAION), the practicality bar is low: Carlini et al. (2024) demonstrated that injecting poisoned examples into ~0.01% of the training corpus is feasible for an attacker controlling a handful of expired domains.

Governance relevance is direct and increasingly cited. NIST AI RMF GenAI Profile (NIST AI 600-1) §2.6 'Information Security' names data poisoning. EU AI Act Art. 15 cybersecurity obligations + Art. 55 systemic-risk obligations require protection against 'attempts to alter the use, behaviour or performance of the system' which covers training-time attacks. China's GenAI Measures Art. 7 mandates legal-source training data, which intersects with poisoning resistance. The governance gap: poisoning resistance is hard to verify post-hoc — once a model is trained, distinguishing poisoned-but-undetected from clean is an open problem. For open-data + open-weight foundation models (Pile, RedPajama, Llama series), poisoning resistance must be designed in at curation time.

How to cite this article

APA

Policy Window. (n.d.). Data Poisoning [Wiki article — Concept]. https://policywindow.org/wiki/data-poisoning

CHICAGO

Policy Window. n.d.. "Data Poisoning." Wiki article (Concept). https://policywindow.org/wiki/data-poisoning.

HARVARD

Policy Window (n.d.) 'Data Poisoning', Wiki article — Concept, available at: https://policywindow.org/wiki/data-poisoning.

OSCOLA

Policy Window, 'Data Poisoning' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/data-poisoning> accessed [date].

BIBTEX

```
@misc{policywindow-data-poisoning,
title  = {Data Poisoning},
author = {Policy Window},
year   = {n.d.},
howpublished = {data-poisoning — safety},
url    = {https://policywindow.org/wiki/data-poisoning},
note   = {Primary source: https://arxiv.org/abs/2302.10149}
}
```