AYI NEDJIMI Consultants

ML Supply Chain 2026 : Backdoors HF & Pickle

16 mai 2026 • Mis à jour le 17 mai 2026 • 20 min de lecture • 3385 mots •

Pickle RCE, backdoors HF, BadNets : la chaine ML est en 2026 le maillon f
modeles HF malveillants identifies, defenses Safetensors et sigstore.

À RETENIR

A retenir — ML Supply Chain Attacks

Pickle RCE : ~3000 modeles malveillants identifies sur Hugging Face en 20
via `__reduce__` deserialization.

BadNets (Gu et al., 2017) etendu aux LLM : insertion de triggers dans le tra
deviation comportementale sur input trigger.

Safetensors (Hugging Face, 2023) elimine le risque pickle, devenu standar
pour les nouveaux modeles.

sigstore + **ML-BOM** (SPDX 3.0) emergent co ttestation e
ML.

> Cas reel 2024 : Stable Diffusion checkpoint compromis sur civitai.com, mi
> deploye sur 12000 machines.

La **ml supply chain attack** exploite un constat simple : un modele neural est une s
d'un graphe de calcul, executee a l'inferenece sur la machine cible avec les meme
l'utilisateur. Si le format de serialisation autorise du code arbitraire (Python pickle)
lui-meme contient un backdoor pre-entraine (BadNets), la chaine d'approvisionne
un vecteur d'attaque de premier ordre. En 2026, le Hub Hugging Face heberge >1
dont plusieurs centaines ont ete identifies comme malveillants. La standardisation
sigstore, ML-BOM) progresse mais reste largement inappliquee dans les entrepris
presente la mecanique des attaques (pickle RCE, BadNets, model spoofing), le co
Python, les defenses 2026, et la conformite (AI Act, OWASP MLSecOps, NIST AI R
CISO et data scientists deployant des modeles tiers en 2026, la **ml supply chain a**
maillon faible le plus sous-investi — un seul `torch.load` non audite peut compro
cluster GPU.

---

1. Genese et etat de l'art

Trois etapes historiques :

> **Pickle deserialization (2018-2024)** — format de facto pour PyTorch (.pt, .pth, .
> Keras (.h5), scikit-learn (.pkl). Le pickle Python autorise l'execution de code arb
> `__reduce__` — faille connue depuis 1997 (Pickle docs, Python). En ML, le risqu
> avec la popularisation de Hugging Face en 2020.

Réponse sous 24h

Devis gratuit →