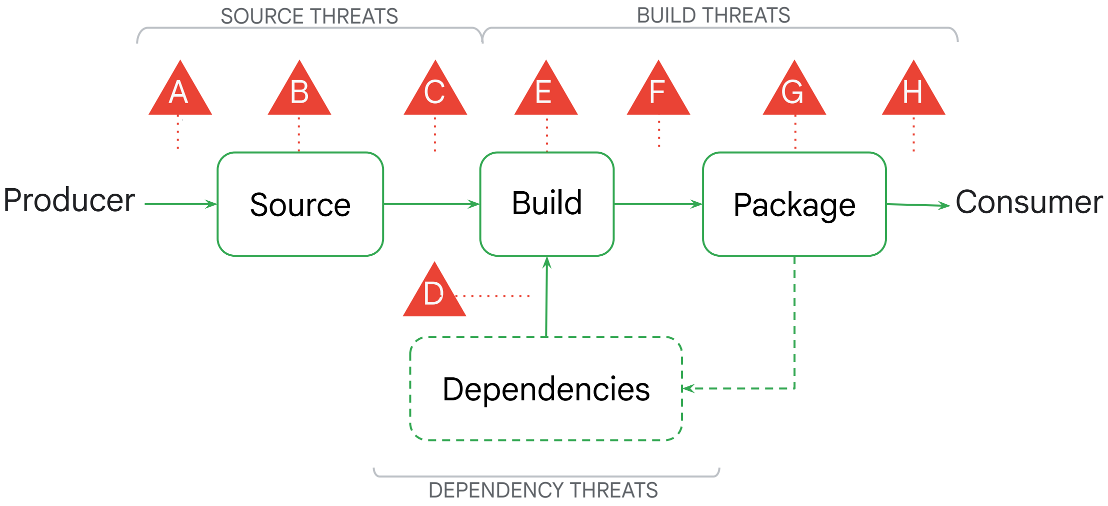
Machine learning models have a supply chain problem

Sarah Meiklejohn, Hayden Blauzvern, Mihai Maruseac, Spencer Schrock, Laurent Simon, and Ilia Shumailov

Security and Privacy Research

Open-source software supply chain



SOURCE THREATS

A Submit unauthorized change

B Compromise source repo

C Build from modified source

DEPENDENCY THREATS

D Use compromised dependency

BUILD THREATS

E Compromise build process

F Upload modified package

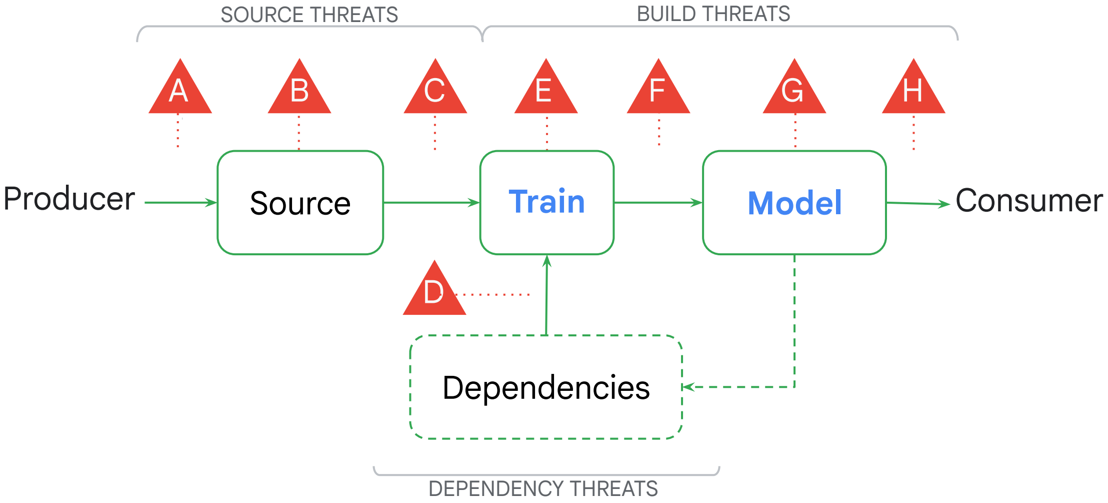
G Compromise package registry

H Use compromised package

Google

Security and Privacy Research

Open ML model supply chain



SOURCE THREATS

A Submit unauthorized change

B Compromise source repo

C **Train** from modified source

DEPENDENCY THREATS

D Use compromised dependency

BUILD THREATS

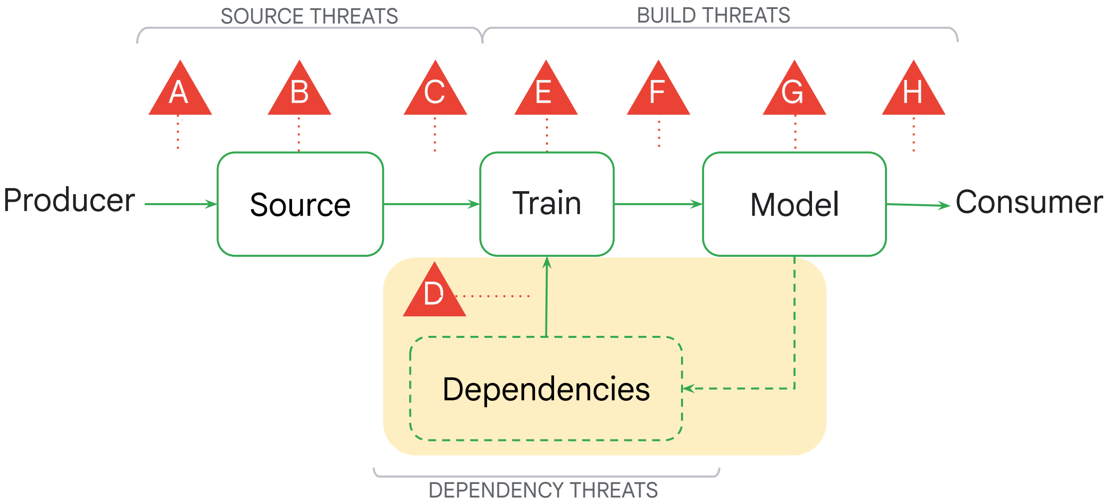
E Compromise **training** process

F Upload modified **model**

G Compromise **model hub**

H Use compromised **model**

Open ML model supply chain



SOURCE THREATS

A Submit unauthorized change

B Compromise source repo

C Train from modified source

DEPENDENCY THREATS

D Use compromised dependency

BUILD THREATS

E Compromise training process

F Upload modified model

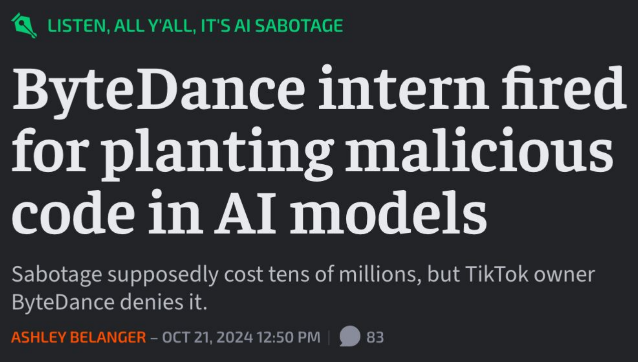
G Compromise model hub

H Use compromised model

The ML community has extensively explored attacks using compromised training data...

Google

Security and Privacy Research

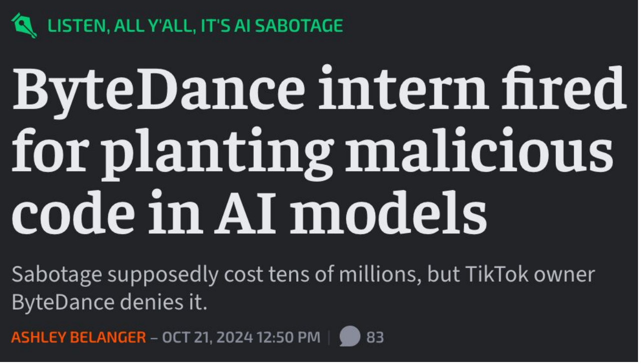


LISTEN, ALL Y'ALL, IT'S AI SABOTAGE

ByteDance intern fired for planting malicious code in AI models

Sabotage supposedly cost tens of millions, but TikTok owner ByteDance denies it.

ASHLEY BELANGER – OCT 21, 2024 12:50 PM | 83

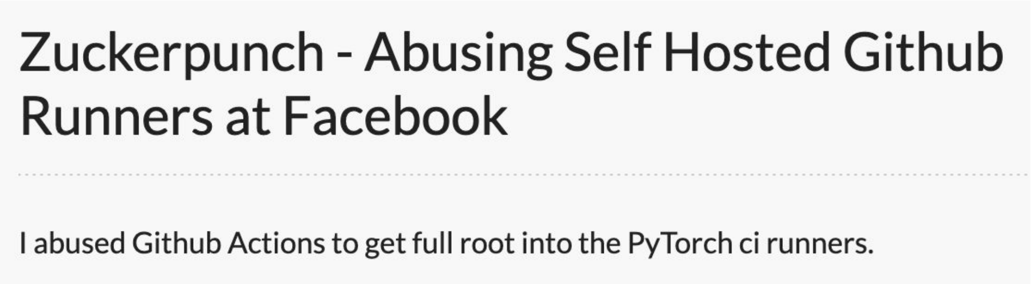


LISTEN, ALL Y'ALL, IT'S AI SABOTAGE

ByteDance intern fired for planting malicious code in AI models

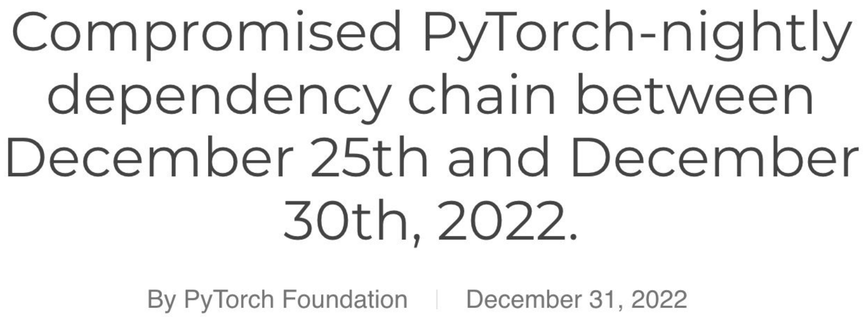
Sabotage supposedly cost tens of millions, but TikTok owner ByteDance denies it.

ASHLEY BELANGER – OCT 21, 2024 12:50 PM | 83



Zuckerpunch - Abusing Self Hosted Github Runners at Facebook

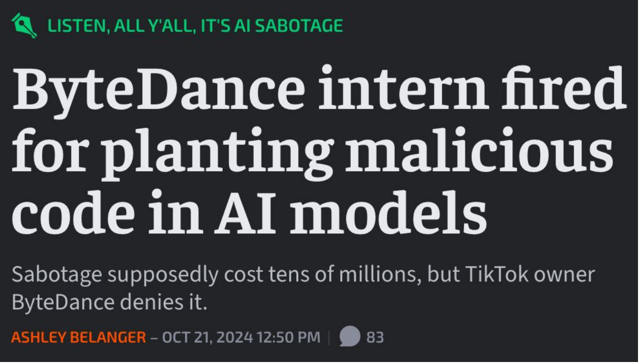
I abused Github Actions to get full root into the PyTorch ci runners.



Compromised PyTorch-nightly dependency chain between December 25th and December 30th, 2022.

By PyTorch Foundation | December 31, 2022

Security and Privacy Research

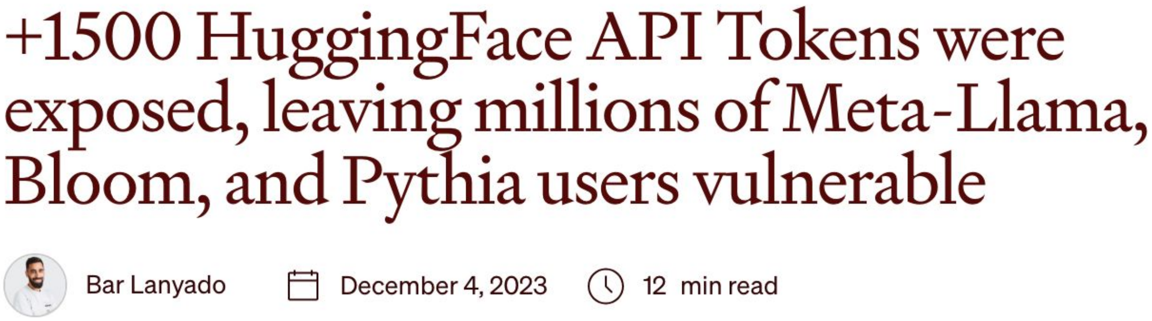


LISTEN, ALL Y'ALL, IT'S AI SABOTAGE

ByteDance intern fired for planting malicious code in AI models

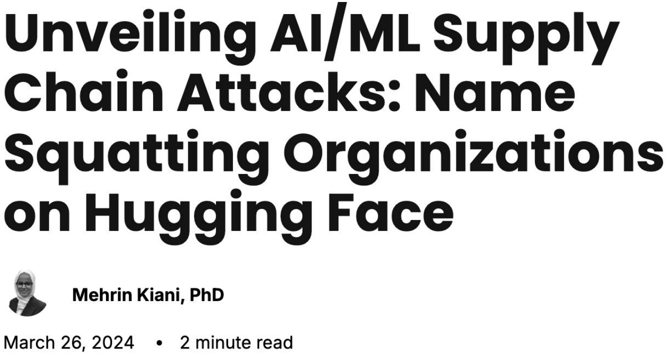
Sabotage supposedly cost tens of millions, but TikTok owner ByteDance denies it.

ASHLEY BELANGER – OCT 21, 2024 12:50 PM | 83



+1500 HuggingFace API Tokens were exposed, leaving millions of Meta-Llama, Bloom, and Pythia users vulnerable

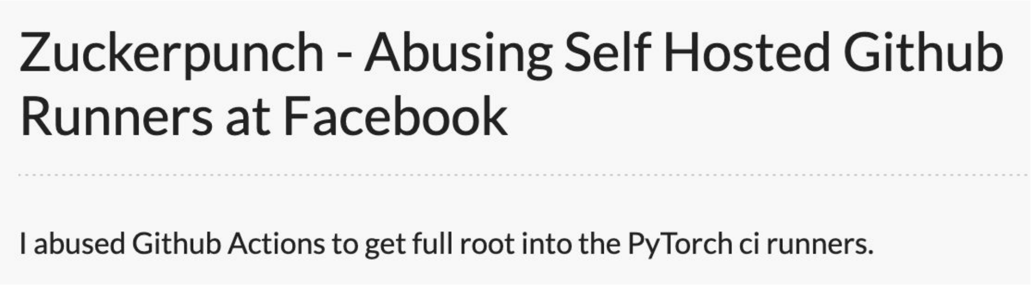
Bar Lanyado December 4, 2023 12 min read



Unveiling AI/ML Supply Chain Attacks: Name Squatting Organizations on Hugging Face

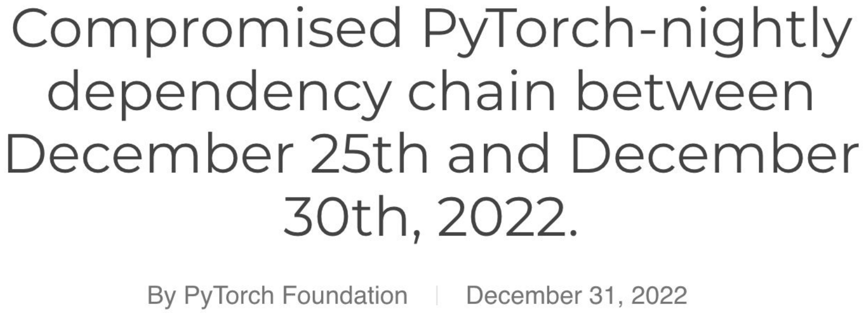
Mehrin Kiani, PhD

March 26, 2024 • 2 minute read



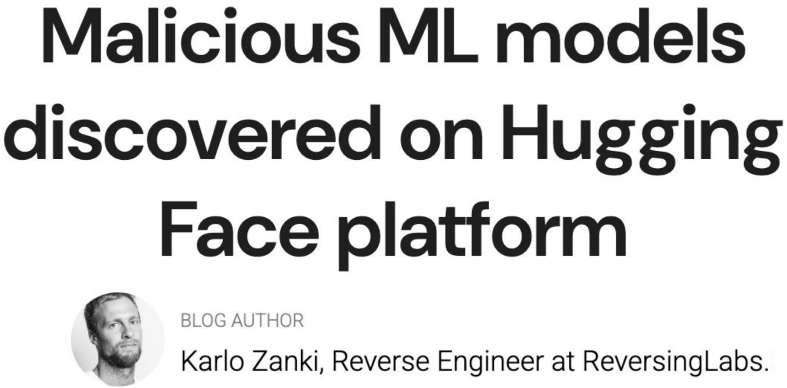
Zuckerpunch - Abusing Self Hosted Github Runners at Facebook

I abused Github Actions to get full root into the PyTorch ci runners.



Compromised PyTorch-nightly dependency chain between December 25th and December 30th, 2022.

By PyTorch Foundation | December 31, 2022

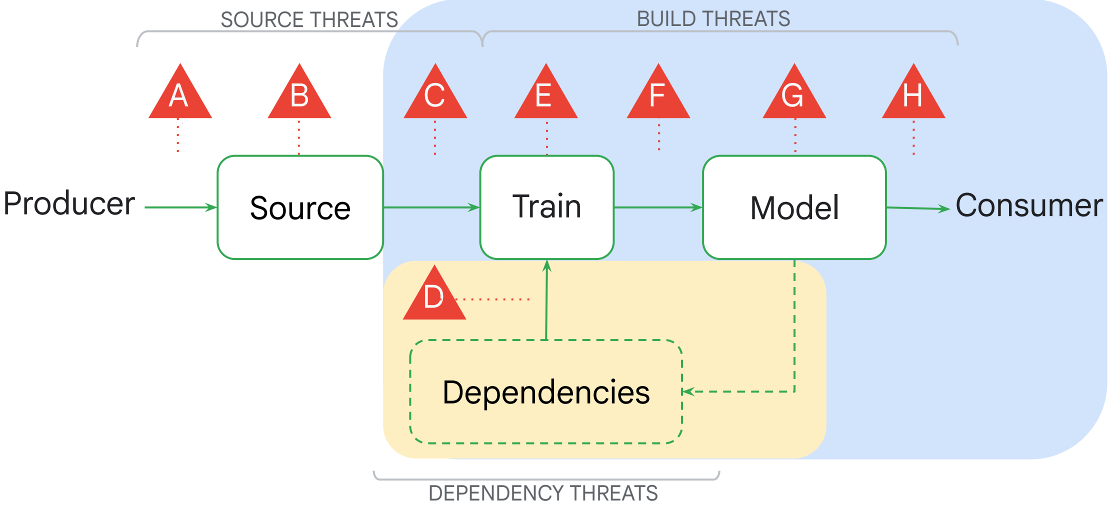


Malicious ML models discovered on Hugging Face platform

BLOG AUTHOR

Karlo Zanki, Reverse Engineer at ReversingLabs.

Open ML model supply chain



...and our paper shows real and potential attacks on a much wider surface!

The ML community has extensively explored attacks using compromised training data...

SOURCE THREATS

A Submit unauthorized change

B Compromise source repo

C Train from modified source

DEPENDENCY THREATS

D Use compromised dependency

BUILD THREATS

E Compromise training process

F Upload modified model

G Compromise model hub

H Use compromised model

Our results

1. **An exploration of attacks on open ML model supply chains**

   They are happening and there is a lot more we could be doing defensively!

2. **Model signing is a first step towards tamper-evident models**

   A performant solution that provides a basic notion of integrity

3. **Cryptographic techniques can be used for dataset transparency**

   These make it possible to prove (non-)inclusion in training data

Thank you!

s.meiklejohn@ucl.ac.uk

Google

Security and Privacy Research