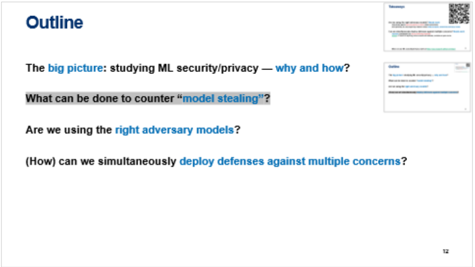
Meta concerns in ML security/privacy

N. Asokan
https://asokan.org/asokan/
@asokan.org @nasokan

(Joint work with Vasisht Duddu, Jian Liu, Sebastian Szyller, Asim Waheed, Rui Zhang)

Outline



The big picture: studying ML security/privacy — why and how?

What can be done to counter “model stealing”?

Are we using the right adversary models?

(How) can we simultaneously deploy defenses against multiple concerns?

Outline

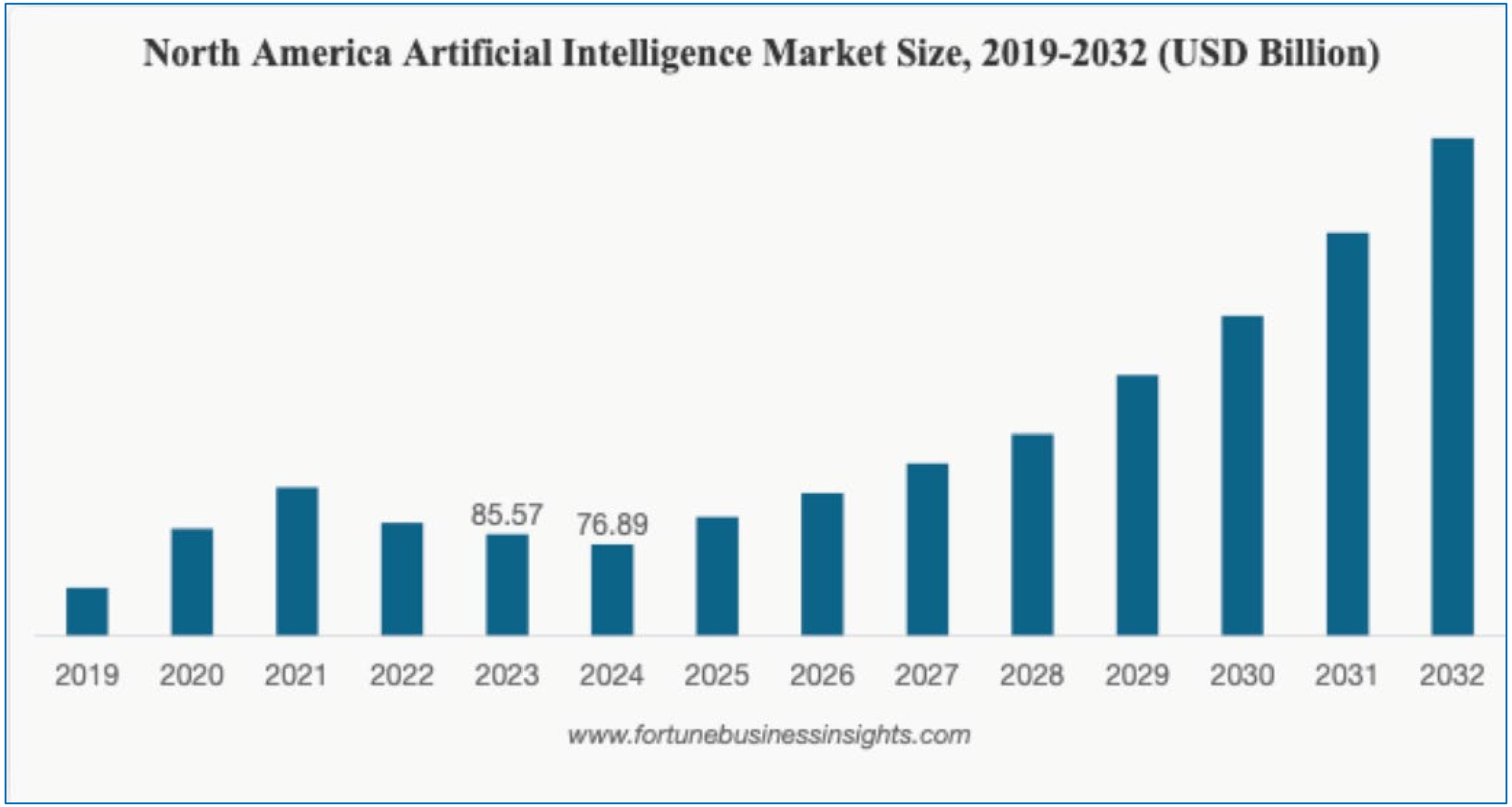
The big picture: studying ML security/privacy — why and how?

What can be done to counter “model stealing”?

Are we using the right adversary models?

(How) can we simultaneously deploy defenses against multiple concerns?

AI will be pervasive



https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-market-100114

Forbes

7,109 views | Oct 18, 2019, 01:56pm EDT

How Artifical Intelligence Is Advancing Precision Medicine

Nicole Martin Former Contributor

AI & Big Data

I write about digital marketing, data and privacy concerns.

https://www.forbes.com/sites/nicolemartin1/2019/10/18/how-artifical-intelligence-is-advancing-precision-medicine/#2f720a79a4d5

VICE NEWSLETTERS

Tech

Dozens of Cities Have Secretly Experimented With Predictive Policing Software

By Caroline Haskins **February 6, 2019, 11:00am**

https://www.vice.com/en_us
experimented-with-predictiv

Forbes Sign In

How AI Is Uprooting Recruiting

By Falon Fatemi , Contributor. Forbes Contributor covering the future of...

Follow Author

Oct 31, 2019, 02:42pm EDT

https://www.forbes.com/sites/falonfatemi/2019/10/31/how-ai-is-uprooting-recruiting/

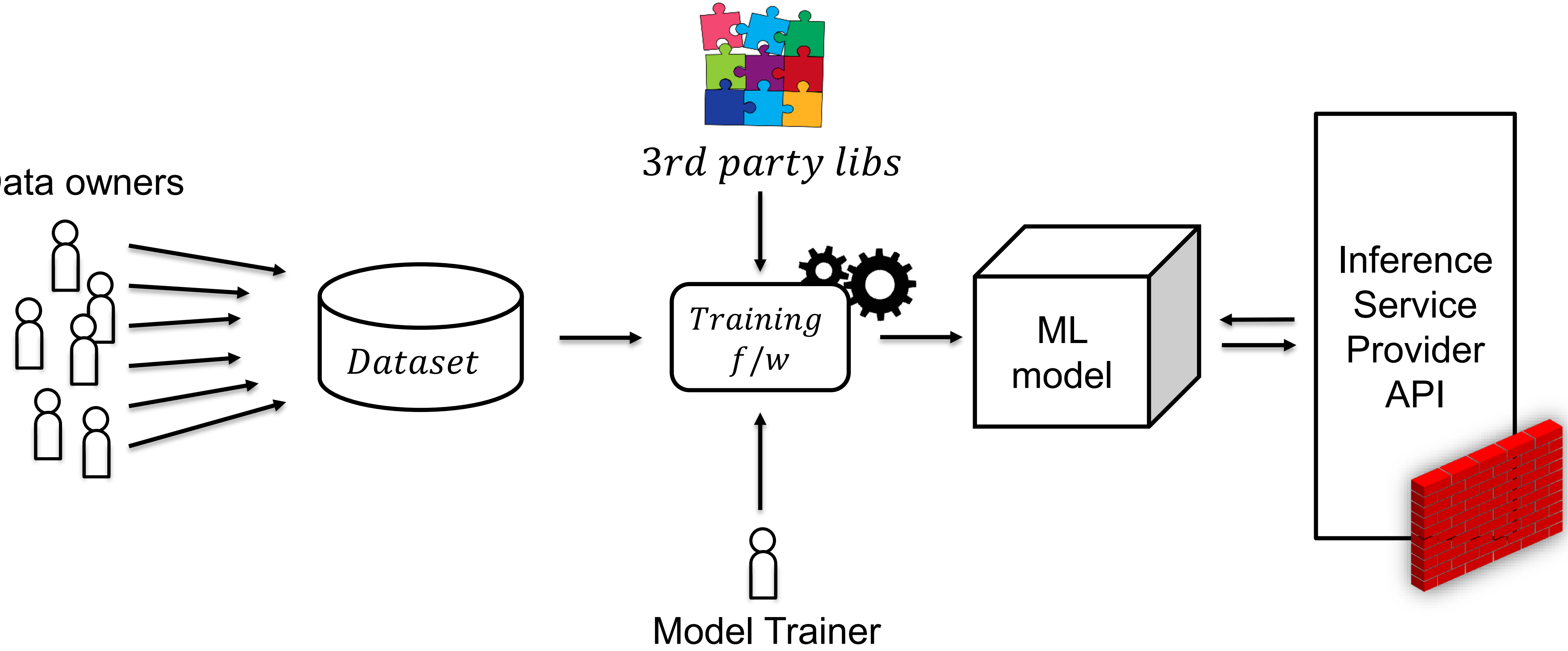
PART OF A ZDNET SPECIAL FEATURE: CYBERSECURITY: LET'S GET TACTICAL

AI is changing everything about cybersecurity for better and for worse. Here's what you nee to know

Artificial intelligence and machine learning tools could go a long way to helping to fight cybercrime. But these technologies aren't a silver bullet, and could also be exploited by malicious hackers.

https://www.zdnet.com/article/ai-is-changing-everything-about-cybersecurity-for-better-and-for-worse-heres-what-you-need-to

Machine Learning pipeline



Where is the adversary? What is its target?

Compromised input – Model integrity

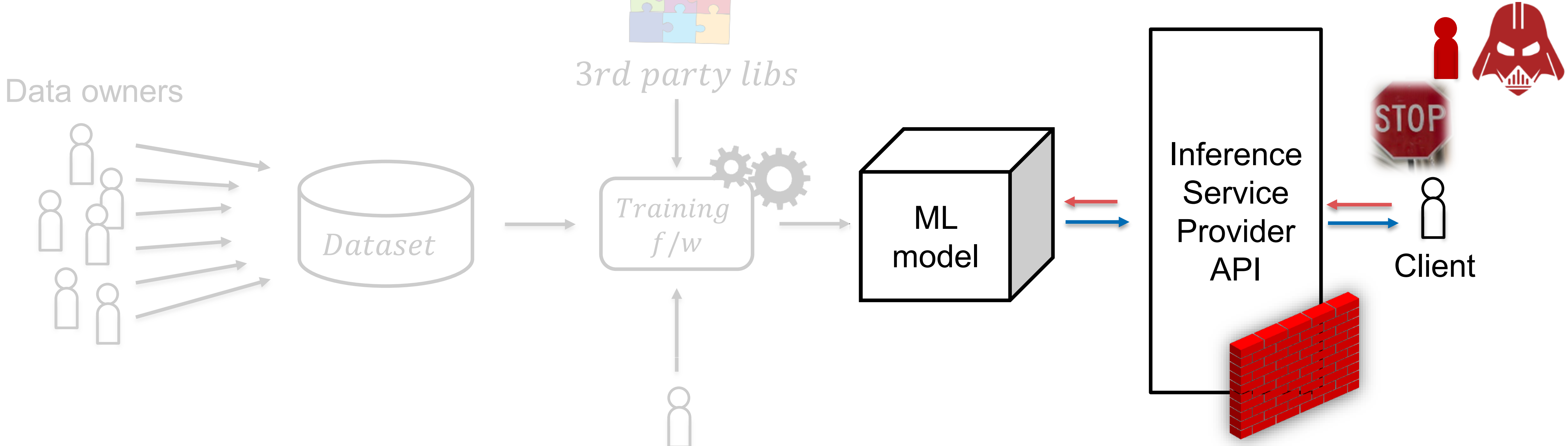
WIRED

WILL KNIGHT SECURITY AUG 1, 2023 7:00 AM

A New Attack Impacts Major AI Chatbots—and No One Knows How to Stop It

Researchers found a simple way to make ChatGPT, Bard, and other chatbots misbehave, proving that AI is hard to tame.

https://www.wired.com/story/ai-adversarial-attacks/



Evade model

Madry et al. – *Towards Deep Learning Models Resistant to Adversarial Attacks*, ICLR '18 (https://arxiv.org/abs/1706.06083)
Carlini & Wagner. – *Towards Evaluating the Robustness of Neural Networks*, IEEE S&P '17 (https://arxiv.org/abs/1608.04644)

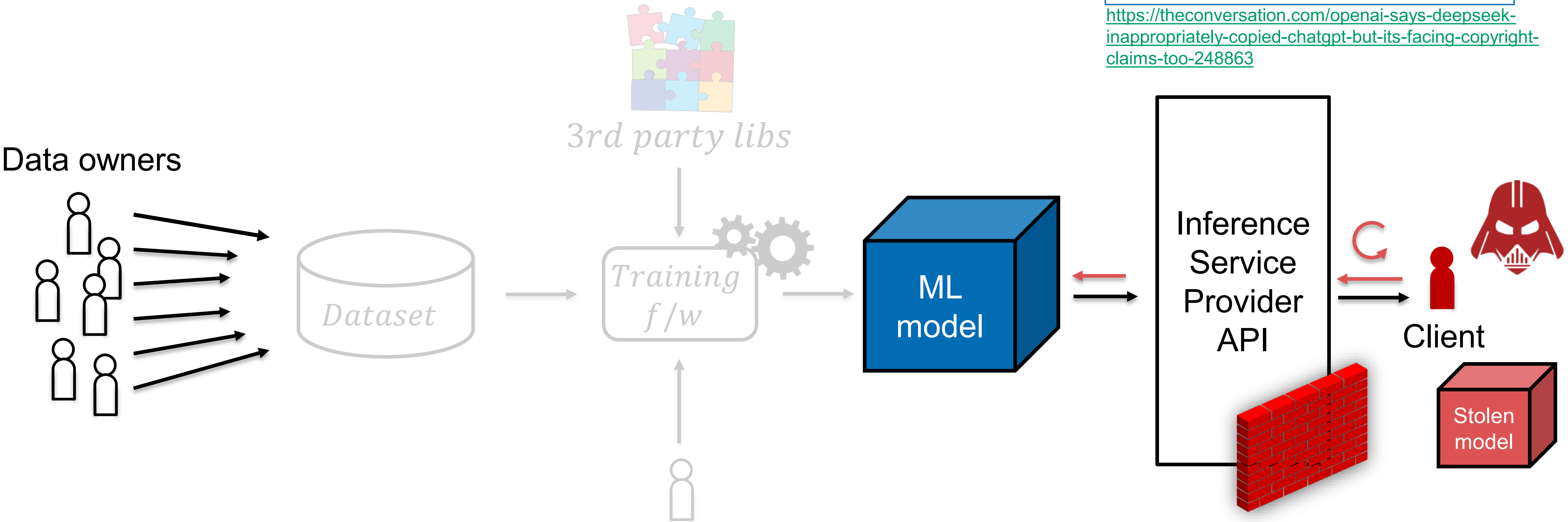
Malicious client – Model confidentiality


OpenAI says DeepSeek ‘inappropriately’ copied ChatGPT – but it’s facing copyright claims too

Published: February 4, 2025 2.10pm EST

Lea Frermann, Shaanan Cohney, *The University of Melbourne*

https://theconversation.com/openai-says-deepseek-inappropriately-copied-chatgpt-but-its-facing-copyright-claims-too-248863



Extract/“steal” model

Tramer et al. – *Stealing ML models via prediction APIs*, Usenix SEC ‘16 (https://arxiv.org/abs/1609.02943)
Juuti et al. – *PRADA: Protecting against DNN Model Stealing Attacks*, Euro S&P ‘19 (https://arxiv.org/abs/1805.02628)
Carlini et al. – *Stealing part of a production language model*, ICML ‘24 (https://arxiv.org/abs/2403.06634)

Towards trustworthy AI

Secure, privacy-preserving, …

TABLE V
TOP ATTACK

Which attack would affect your org the most?	*Distribution*
Poisoning (e.g: [21])	10
Model Stealing (e.g: [22])	6
Model Inversion (e.g: [23])	4
Backdoored ML (e.g: [24])	4
Membership Inference (e.g: [25])	3
Adversarial Examples (e.g: [26])	2
Reprogramming ML System (e.g: [27])	0
Adversarial Example in Physical Domain (e.g: [5])	0
Malicious ML provider recovering training data (e.g: [28])	0
Attacking the ML supply chain (e.g: [24])	0
Exploit Software Dependencies (e.g: [29])	0

Kumar et al. – *Adversarial Machine Learning – Industry Perspectives*, IEEE SPW ‘20 (https://arxiv.org/abs/2002.05646)

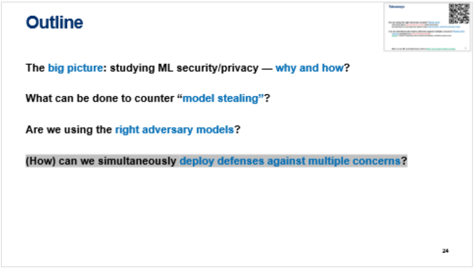
Outline

The big picture: studying ML security/privacy — why and how?

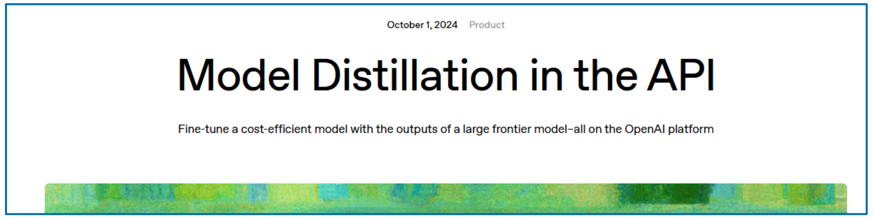
What can be done to counter “model stealing”?

Are we using the right adversary models?

(How) can we simultaneously deploy defenses against multiple concerns?



Defending against model stealing


October 1, 2024 Product

Model Distillation in the API

Fine-tune a cost-efficient model with the outputs of a large frontier model–all on the OpenAI platform

https://openai.com/index/api-model-distillation/


The New York Times

Artificial Intelligence › OpenAI and Trump A Look at OpenAI's Operator What Is DeepSeek? DeepSeek's Rise Quiz

OpenAI Says DeepSeek May Have Improperly Harvested Its Data

The San Francisco start-up claims that its Chinese rival may have used data generated by OpenAI technologies to build new systems.

https://www.nytimes.com/2025/01/29/technology/openai-deepseek-data-harvest.html


"We are aware of and reviewing indications that DeepSeek may have inappropriately distilled our models, and will share information as we know more," the spokesperson said, adding that the company was not accusing DeepSeek of a security breach.

Distillation is often prohibited in LLMs' terms of service, but is common in the industry.

We can try to:

- prevent (or slow down[1]) model extraction, or
- detect[2] it

But current solutions are not effective

Model derivation may even become a desirable business model

Deter unauthorized model ownership via model ownership resolution (MOR):

- watermarking
- fingerprinting

[1] Dziedzic et al. – *Increasing the Cost of Model Extraction with Calibrated Proof of Work*, ICLR '22 (https://openreview.net/pdf?id=EAy7C1cgE1L)
[2] Atli et al. – *Extraction of Complex DNN Models: Real Threat or Boogeyman?*, AAAI-EDSML '20 (https://arxiv.org/abs/1910.05429)

Watermarking

Embed watermark while training (potentially) victim model[1]

- Choose incorrect labels for a set of samples (watermark set, WM)
- Cannot resist model extraction

Embed watermark at the inference API[2]

- Use a mapping function to decide when to return incorrect predictions for queries
- Finding suitable mapping functions is difficult

Watermarking schemes tend to be not robust[3] and reduce utility

[1] Yadi et al. – *Watermarking Deep Neural Networks by Backdooring,* Usenix SEC '18 https://www.usenix.org/node/217594
[2] Szyller et. al. – *DAWN: Dynamic Adversarial Watermarking of Neural Networks,* ACM MM '21 (https://arxiv.org/abs/1906.00830)
[3] Lukas et al. – *SoK: How Robust is Image Classification Deep Neural Network Watermarking*? IEEE S&P '22 (https://arxiv.org/abs/2108.04974)

Fingerprinting

Conferrable adversarial examples[1]

- Distinguish between conferrable adversarial examples vs. other transferable ones
- Computationally expensive

Dataset inference[2]

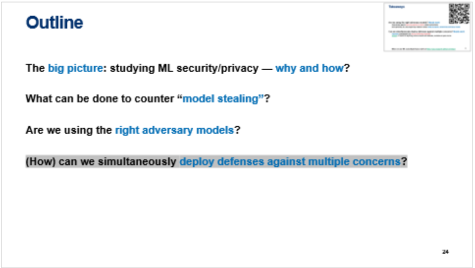
- Distinguish between models trained with different datasets
- Susceptible to false positives/negatives under certain conditions[3]

GrOVe[4]

- Use GNN embeddings as fingerprints (for GNN models)
- Effective against high-fidelity extraction[5] but likely not against low-fidelity extraction

[1] Lukas et al. – *Deep Neural Network Fingerprinting by Conferrable Adversarial Examples,* ICLR '21 (https://openreview.net/forum?id=VqzVhqxkjH1)
[2] Maini et al. – *Dataset Inference Ownership Resolution in Machine Learning*, ICLR '21 (https://openreview.net/pdf?id=hvdKKV2yt7T)
[3] Szyller et al. – *On the Robustness of Dataset Inference*, TMLR '23 (https://arxiv.org/abs/2210.13631)
[4] Waheed et al. – *GrOVe: Ownership Verification of Graph Neural Networks using Embeddings,* IEEE S&P '24 (https://arxiv.org/abs/2304.08566)
[5] Shen et al. – *Model Stealing Attacks Against Inductive Graph Neural Networks,* IEEE S&P '22 (https://arxiv.org/abs/2112.08331)

Outline



The big picture: studying ML security/privacy — why and how?

What can be done to counter "model stealing"?

Are we using the right adversary models?

(How) can we simultaneously deploy defenses against multiple concerns?

Robustness of model ownership resolution schemes

Model ownership resolution (MOR) must be robust against adversaries

Malicious suspect:
- tries to evade verification (e.g., pruning, fine-tuning, noising)

Malicious accuser:
- tries to frame an independent model owner
- (secure) timestamping (watermark/fingerprint and model) is the only defense in prior work

So far, research has focused on robustness against malicious suspects

False claims against MORs

We show how malicious accusers can make false claims against independent models:

- adversary deviates from watermark/fingerprint generation procedure
  - E.g., via transferrable adversarial examples
- but still subject to specified verification procedure

Our contributions:

- formalize the notion of false claims against MORs
- provide a generalization of MORs
- demonstrate effective false claim attacks
- discuss potential countermeasures

Zhang et al. – *False Claims Against Model Ownership Resolution*, Usenix SEC ‘24 (https://arxiv.org/abs/2304.06607)

Watermarking by backdooring[1]

Watermark generation:

- choose some out-of-distribution samples as watermark
  - assigned with incorrect labels
- train using the watermark alongside normal training data (or fine tune)
  - model memorizes watermark
- obtain timestamp on commitment of model and watermark

Watermark verification:

- query suspect model using watermark
- compare predictions to the assigned (incorrect) labels:
  - many matching / high WM accuracy → stolen
  - a few matching / low WM accuracy → not stolen
- check commitment and timestamp

[1] Adi et al. – *Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring*, Usenix SEC 2018 (https://arxiv.org/abs/1802.04633)

Watermarking by backdooring[1]: false claim[2]

Watermark generation:

- choose some out-of-distribution samples as watermark
  - assigned with incorrect labels
- train using the watermark alongside your normal training data (or fine tune)
  - model memorizes watermark
- obtain timestamp on commitment of model and watermark

Watermark verification:

- query suspect model using watermark
- compare predictions to the assigned (incorrect) labels:
  - many matching / high WM accuracy → stolen
  - a few matching / low WM accuracy → not stolen
- check commitment and timestamp

[1] Adi et al. – *Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring*, Usenix SEC 2018 (https://arxiv.org/abs/1802.04633)
[2] Zhang et al. – *False Claims Against Model Ownership Resolution*, Usenix SEC ‘24 (https://arxiv.org/abs/2304.06607)

Watermarking by backdooring[1]: false claim[2]

False watermark generation:
- choose some out-of-distribution samples
- perturb them to craft transferable adversarial examples → false watermark
- obtain timestamp on commitment of model and false watermark

Watermark verification:
- query suspect model using watermark
- compare predictions to the assigned (incorrect) labels:
  - many matching / high WM accuracy -> stolen
  - a few matching / low WM accuracy > not stolen
- check commitment and timestamp

[1] Adi et al. – *Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring*, Usenix SEC 2018 (https://arxiv.org/abs/1802.04633)
[2] Zhang et al. – *False Claims Against Model Ownership Resolution*, Usenix SEC '24 (https://arxiv.org/abs/2304.06607)

Mitigating false claims against MORs

Judge generates watermarks/fingerprints: bottleneck

Judge verifies watermarks/fingerprints were generated correctly: expensive

Train models with transferable adversarial examples: accuracy loss

Zhang et al. – *False Claims Against Model Ownership Resolution*, Usenix SEC '24 (https://arxiv.org/abs/2304.06607)

The Meta Concern: sensible adversary models

Identify potential adversaries and their goals systematically

Identify adversary's knowledge and capabilities:

- **Data access**:
  - vis-à-vis target's *training data* (overlap/distribution/domain? natural/synthetic?)
  - vis-à-vis target's *inferences*
- Target **model access**: white-box/black-box/grey-box?
- **Adversary type**: honest-but-curious vs. malicious
- **Interaction type**: zero-shot/one-shot/query-budget?, adaptive?

Avoid sloppy terminology!

- "adversarial attacks" → there are no benign attacks!
- "adaptive adversaries" → cf. Kerchoff's principle!
- …

Outline

The big picture: studying ML security/privacy — why and how?

What can be done to counter “model stealing”?

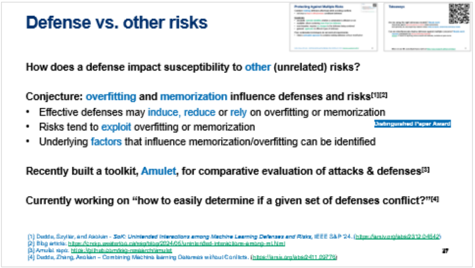
Are we using the right adversary models?

(How) can we simultaneously deploy defenses against multiple concerns?

Unintended interactions

Prior work explored defenses to mitigate specific risks

- Defenses typically evaluated only vs. specific risks they protect against



But practitioners need to deploy multiple defenses simultaneously

- Can two defenses interact negatively with each other?
- Does a defense exacerbate or ameliorate some other (unrelated) risk?

Szyller and Asokan – *Conflicting Interactions Among Protections Mechanisms for Machine Learning Models*, AAAI '23 (https://arxiv.org/abs/2207.01991)

Ownership resolution vs. other security/privacy concerns

There are considerations other than model ownership resolution:

- model evasion (defense: adversarial training)
- training data reconstruction (defense: differential privacy)
- membership inference (defense: regularization, early stopping)
- model poisoning (defense: regularization, outlier/anomaly detection)
- …

How do ownership resolution schemes interact with the other defenses?

We investigated pairwise interactions of:

model watermarking data watermarking fingerprinting	**WITH**	differential privacy adversarial training

Szyller and Asokan – *Conflicting Interactions Among Protections Mechanisms for Machine Learning Models*, AAAI ‘23 (https://arxiv.org/abs/2207.01991)

Ownership resolution vs. other security/privacy concerns

If two techniques A and B in combination result in too high a drop in

- model accuracy (ϕ_{ACC}) or
- metric for A (ϕ_A) or
- metric for B (ϕ_B)

then A and B are in conflict

Defense	Dataset	Defense	
		DP	ADV. TR.
WM	MNIST	ϕ_{ACC} ϕ_{WM}	ϕ_{ACC} ϕ_{WM} ϕ_{ADV}
	FMNIST	ϕ_{ACC} ϕ_{WM}	ϕ_{ACC} ϕ_{WM} ϕ_{ADV}
	CIFAR10	ϕ_{ACC} ϕ_{WM}	ϕ_{ACC} ϕ_{WM} ϕ_{ADV}
RAD-DATA	MNIST	ϕ_{ACC} $\phi_{RAD\text{-}DATA}$	ϕ_{ACC} $\phi_{RAD\text{-}DATA}$ ϕ_{ADV}
	FMNIST	ϕ_{ACC} $\phi_{RAD\text{-}DATA}$	ϕ_{ACC} $\phi_{RAD\text{-}DATA}$ ϕ_{ADV}
	CIFAR10	ϕ_{ACC} $\phi_{RAD\text{-}DATA}$	ϕ_{ACC} $\phi_{RAD\text{-}DATA}$ ϕ_{ADV}
DI	MNIST	ϕ_{ACC} ϕ_{DI}	ϕ_{ACC} ϕ_{DI} ϕ_{ADV}
	FMNIST	ϕ_{ACC} ϕ_{DI}	ϕ_{ACC} ϕ_{DI} ϕ_{ADV}
	CIFAR10	ϕ_{ACC} ϕ_{DI}	ϕ_{ACC} ϕ_{DI} ϕ_{ADV}

Szyller and Asokan – *Conflicting Interactions Among Protections Mechanisms for Machine Learning Models*, AAAI '23 (https://arxiv.org/abs/2207.01991)

Interaction between ML defenses

Property	Adversarial Training	Differential Privacy	Membership Inference	Oblivious Training	Model/Gradient Inversion	Model Poisoning	Model Watermarking	Model Fingerprinting	Data Watermarking	Explainability	Fairness
Adversarial Training	X	[5]	[9]	?	?	[7]	OURS	OURS	OURS	[11]	?
Differential Privacy		X	[3, 6]	?	?	?	OURS	OURS	OURS	?	[1, 2, 8]
Membership Inference			X	?	?	[10]	?	?	?	?	?
Oblivious Training				X	?	?	?	?	?	?	?
Model/Gradient Inversion					X	?	?	?	?	?	?
Model Poisoning						X	?	?	?	?	?
Model Watermarking							X	?	?	?	?
Model Fingerprinting								X	?	[4]	?
Data Watermarking									X	?	?
Fairness										X	?
Explainability											X

REFERENCES

[1] Hongyan Chang and Reza Shokri. 2021. On the Privacy Risks of Algorithmic Fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS P)*. 292–303. https://doi.org/10.1109/EuroSP51992.2021.00028

[2] Victoria Cheng, Vinith M. Suriyakumar, Natalie Dullerud, Shalmali Joshi, and Marzyeh Ghassemi. 2021. Can You Fake It Until You Make It? Impacts of Differentially Private Synthetic Data on Downstream Classification Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 149–160. https://doi.org/10.1145/3442188.3445879

[3] Thomas Humphries, Simon Oya, Lindsey Tulloch, Matthew Rafuse, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. 2020. Investigating Membership Inference Attacks under Data Dependencies. https://doi.org/10.48550/ARXIV.2010.12112

[4] Hengrui Jia, Hongyu Chen, Jonas Guan, Ali Shahin Shamsabadi, and Nicolas Papernot. 2022. A Zest of LIME: Towards Architecture-Independent Model Distances. In *International Conference on Learning Representations*. https://openreview.net/forum?id=OUz_9TiTv9j

[5] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified Robustness to Adversarial Examples with Differential Privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. 656–672. https://doi.org/10.1109/SP.2019.00044

[6] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. 2021. Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning. In *2021 IEEE Symposium on Security and Privacy (SP)*. 866–882. https://doi.org/10.1109/SP40001.2021.00069

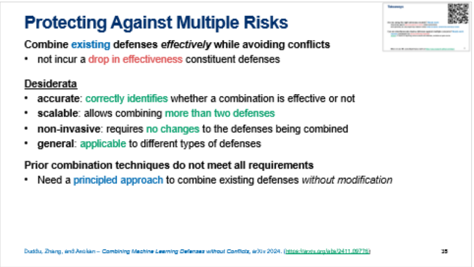
[7] Ren Pang, Hua Shen, Xinyang Zhang, Shouling Ji, Yevgeniy Vorobeychik, Xiapu Luo, Alex Liu, and Ting Wang. 2020. *A Tale of Evil Twins: Adversarial Inputs versus Poisoned Models*. Association for Computing Machinery, New York, NY, USA, 85–99. https://doi.org/10.1145/3372297.3417253

[8] Adam Pearce. 2022. Can a Model Be Differentially Private and Fair? https://pair.withgoogle.com/explorables/private-and-fair/. Online; accessed 7 April 2022.

[9] Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Privacy Risks of Securing Machine Learning Models against Adversarial Examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. Association for Computing Machinery, 241–257. https://doi.org/10.1145/3319535.3354211

[10] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. 2022. Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. https://doi.org/10.48550/ARXIV.2204.00032

[11] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness May Be at Odds with Accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. https://openreview.net/forum?id=SyxAb30cY7


Protecting Against Multiple Risks

Combine existing defenses *effectively* while avoiding conflicts
- not incur a drop in effectiveness constituent defenses

Desiderata
- accurate: correctly identifies whether a combination is effective or not
- scalable: allows combining more than two defenses
- non-invasive: requires no changes to the defenses being combined
- general: applicable to different types of defenses

Prior combination techniques do not meet all requirements
- Need a principled approach to combine existing defenses *without modification*


Takeaways

Are we using the right adversary models? Needs work

Robustness against false accusations in MORs needs improvement

More generally, ML security/privacy research needs widely accepted, streamlined adversary models

Can we simultaneously deploy defenses against multiple concerns? Needs work

Important consideration but not yet sufficiently explored

Amulet: A Toolkit for exploring various attacks and defenses, available as open source

More on our ML security/privacy work at https://ssg-research.github.io/mlsec/

Defense vs. other risks

How does a defense impact susceptibility to other (unrelated) risks?

Conjecture: overfitting and memorization influence defenses and risks[1][2]
- Effective defenses may induce, reduce or rely on overfitting or memorization
- Risks tend to exploit overfitting or memorization
- Underlying factors that influence memorization/overfitting can be identified

Distinguished Paper Award

Recently built a toolkit, Amulet, for comparative evaluation of attacks & defenses[3]

Currently working on "how to easily determine if a given set of defenses conflict?"[4]

[1] Duddu, Szyller, and Asokan - *SoK: Unintended Interactions among Machine Learning Defenses and Risks,* IEEE S&P '24. (https://arxiv.org/abs/2312.04542)
[2] Blog article: https://crysp.uwaterloo.ca/ssg/blog/2024/05/unintended-interactions-among-ml.html
[3] Amulet repo: https://github.com/ssg-research/amulet
[4] Duddu, Zhang, Asokan – Combining Machine learning Defenses without Conflicts. (https://arxiv.org/abs/2411.09776)

Factors influencing overfitting and memorization

O1 Curvature smoothness of the objective function

O2 Distinguishability across datasets (O2.1), subgroups (O2.2), and models (O2.3)

O3 Distance of training data to decision boundary

D1 Size of training data

D2 Tail length of distribution

D3 Number of attributes

D4 Priority of learning stable attributes

M1 Model capacity

Blog article: https://blog.ssg.aalto.fi/2024/05/unintended-interactions-among-ml.html

Duddu, Szyller, and Asokan - *SoK: Unintended Interactions among Machine Learning Defenses and Risks,* IEEE S&P '24. (https://arxiv.org/abs/2312.04542)

Framework: systematizing defenses vs. other risks

Effectiveness of defense <d> correlates with a change in factor <f>

Change in <f> correlates with change in susceptibility to risk <r>

- ↑: positive correlation; ↓: negative correlation

Identify <f> impacted by <d>, and <r> influenced by changes in <f>

Defences (<↑ or ↓>, `<f>`)	Risks (<↑ or ↓>, `<f>`)
RD1 (Adversarial Training): • D1 ↑, $\|\mathcal{D}_{tr}\|$ [161] • D2 ↓, tail length [71], [16] • D4 ↑, priority for learning stable attributes [161] • O1 ↑, curvature smoothness [102] • O2.1 ↑, distinguishability in data records inside and outside $\mathcal{D}_{tr}$ [144] • O3 ↑, distance to boundary for most $\mathcal{D}_{tr}$ data records [176] • M1 ↑, model capacity [102] **RD2 (Outlier Removal):** • D2 ↑, tail length [166] **RD3 (Watermarking):** • D2 ↑, tail length [96] • O2.3 ↓, distinguishability in observables for watermarks between f_θ and f_θ^{der}, but distinct from independent models [3] • M1 ↑, model capacity [3]	**R1 (Evasion):** • D2 ↑, tail length [173], [91] • O1 ↓, curvature smoothness [102] • O3 ↓, distance of $\mathcal{D}_{tr}$ data records to boundary [162] **R2 (Poisoning):** • D2 ↑, tail length [120], [17], [96] • M1 ↑, model capacity [3] **R3 (Unauthorized Model Ownership):** • M1 ↓, model capacity [117], [88] **P1 (Membership Inference):** • D1 ↓, $\|\mathcal{D}_{tr}\|$ [184], [136] • D2 ↑, tail length [25], [24] • D4 ↓, priority for learning stable attributes [103], [155] • O2.1 ↑, distinguishability for data records inside and outside $\mathcal{D}_{tr}$ [136]

Blog article: https://blog.ssg.aalto.fi/2024/05/unintended-interactions-among-ml.html

Duddu, Szyller, and Asokan - *SoK: Unintended Interactions among Machine Learning Defenses and Risks,* IEEE S&P '24. (https://arxiv.org/abs/2312.04542)

Situating prior work in the framework

Risk increases (🔴) or decreases (🟢) or unexplored (⚪) when a defense is effective

Evaluate the influence of factors empirically (●), theoretically (⊙), conjectured (○)

Defenses	Risks		OVFT	Memorization					Both		References
			D1	D2	D3	D4	O1	O2	O3	M1	
RD1 (Adversarial Training)	R1 (Evasion)	🟢		●			●		●	●	[193], [102], [91], [173]
	R2 (Poisoning)	🔴									[170], [153]
	R3 (Unauthorized Model Ownership)	🔴	○								[86] ([95]: 🟢)
	P1 (Membership Inference)	🔴	⊙, ●					1: ●		●	[144], [67]
	P2 (Data Reconstruction)	🔴				○				●	[195], [111]
	P3 (Attribute Inference)	⚪									
	P4 (Distribution Inference)	🔴				○					[148]
	F (Discriminatory Behaviour)	🔴		⊙, ●							[16], [36], [71], [99]
RD2 (Outlier Removal)	R1 (Evasion)	🟢									[59]
	R2 (Poisoning)	🟢									[154]
	R3 (Unauthorized Model Ownership)	⚪									
	P1 (Membership Inference)	🔴		●							[25], [46]
	P2 (Data Reconstruction)	⚪									
	P3 (Attribute Inference)	🔴		●							[78]
	P4 (Distribution Inference)	⚪									
	F (Discriminatory Behaviour)	🔴	●	○							[134]
RD3 (Watermarking)	R1 (Evasion)	⚪									
	R2 (Poisoning)	🔴		○							[133], [3], [194], [93]
	R3 (Unauthorized Model Ownership)	🟢		○				3: ●	●		[152], [3], [98]
	P1 (Membership Inference)	🔴		○				1: ●	●		[157], [33]
	P2 (Data Reconstruction)	🔴		○				1: ●	●		[157]
	P3 (Attribute Inference)	🔴		○				2: ●	●		[157]
	P4 (Distribution Inference)	🔴	⊙, ●	○				1: ●	●	●	[30], [105]

Blog article: https://blog.ssg.aalto.fi/2024/05/unintended-interactions-among-ml.html

Duddu, Szyller, and Asokan - *SoK: Unintended Interactions among Machine Learning Defenses and Risks,* IEEE S&P '24. (https://arxiv.org/abs/2312.04542)

Guideline for conjecturing unintended interactions

For defense <d>, risk <r> and common factor <f>, use pair of arrows that describe how <d> and <r> correspond to <f>

Conjectured interaction for a given <f>:

- If arrows align (↑,↑) or (↓,↓) ➞ <r> increases when <d> is effective (🔴)
- Else for (↑,↓) or (↓,↑) ➞ <r> decreases when <d> is effective (🟢)

Conjectured overall interaction: consider conjectures from all <f>s:

- If all <f> agree, then conjectured overall interaction is unanimous
- Otherwise, prioritize conjecture from dominant <f> (dominance may depend on attack)
- Value of a non-common factor may affect overall interaction

Blog article: https://blog.ssg.aalto.fi/2024/05/unintended-interactions-among-ml.html
Duddu, Szyller, and Asokan - *SoK: Unintended Interactions among Machine Learning Defenses and Risks,* IEEE S&P ‘24. (https://arxiv.org/abs/2312.04542)


Takeaways

Are we using the right adversary models? Needs work

Robustness against false accusations in MORs needs improvement

More generally, ML security/privacy research needs widely accepted, streamlined adversary models

Can we simultaneously deploy defenses against multiple concerns? Needs work

Important consideration but not yet sufficiently explored

Amulet: A Toolkit for exploring various attacks and defenses, available as open source

More on our ML security/privacy work at https://ssg-research.github.io/mlsec/

Protecting Against Multiple Risks

Combine existing defenses *effectively* while avoiding conflicts

- not incur a drop in effectiveness constituent defenses

Desiderata

- **accurate**: correctly identifies whether a combination is effective or not
- **scalable**: allows combining more than two defenses
- **non-invasive**: requires no changes to the defenses being combined
- **general**: applicable to different types of defenses

Prior combination techniques do not meet all requirements

- Need a principled approach to combine existing defenses *without modification*

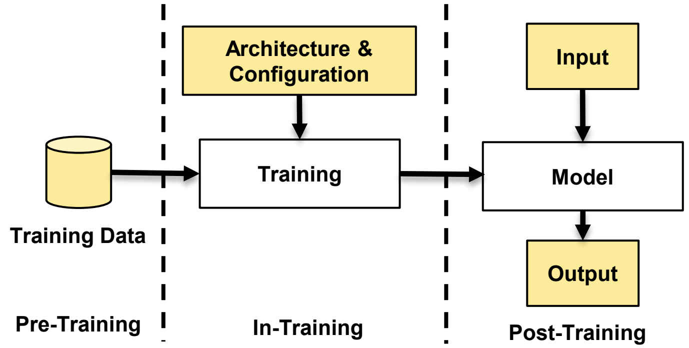
Duddu, Zhang, and Asokan – *Combining Machine Learning Defenses without Conflicts,* arXiv 2024. (https://arxiv.org/abs/2411.09776)

Combining ML Defenses without Conflicts

Intuition: account for reasons underlying conflicts among defenses

For D_1 and D_2 applied in that order, there can be a conflict if

- D_1 uses a risk protected by D_2
- Changes by D_2 overrides changes by D_1

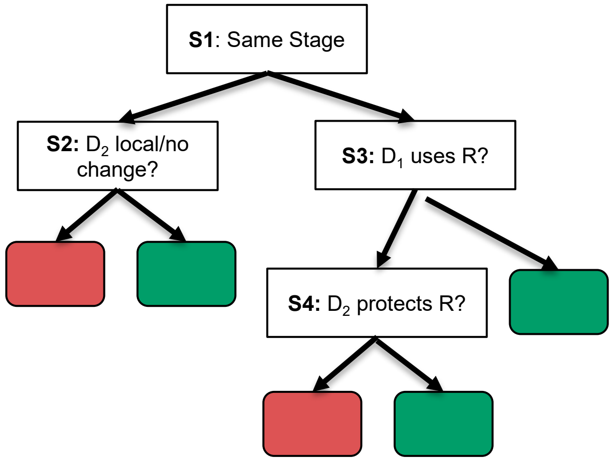


Observation:

- ML defenses operate on one of three stages of ML pipelines

DEF\CON: quickly identify effective combinations

- 90% accuracy on eight combinations from prior work
- 81% in 30 previously unexplored combinations



Duddu, Zhang, and Asokan – *Combining Machine Learning Defenses without Conflicts,* arXiv 2024. (https://arxiv.org/abs/2411.09776)

Amulet: Introduction

Python package to evaluate susceptibility of risks to security, privacy, and fairness

Goals

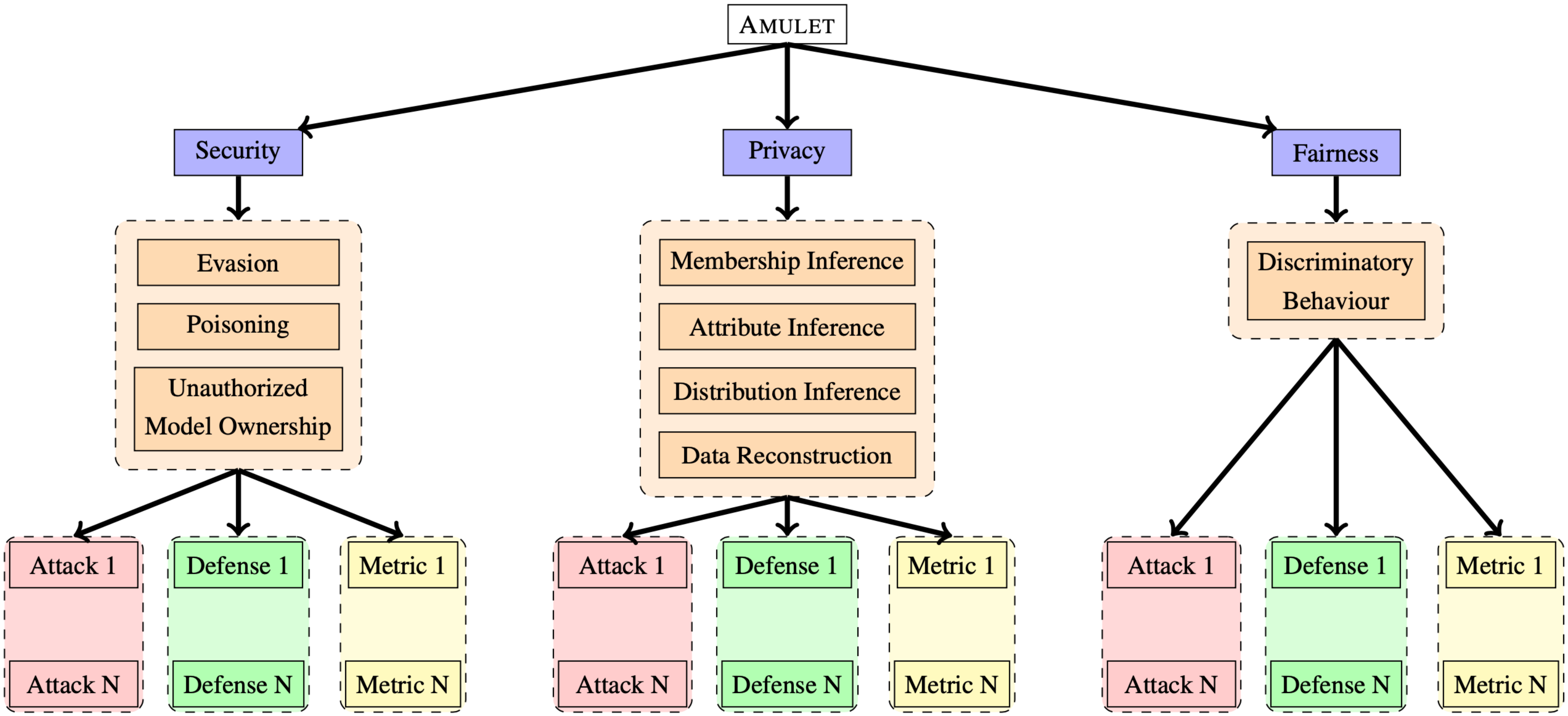
- Evaluate how algorithms designed to reduce one risk may impact another unrelated risk
- Compare different attacks/defenses for a given risk

Desiderata

- **Comprehensive**: Covers most representative attacks/defenses/metrics for different risks
- **Extensible**: Easy to include additional risks, attacks, defenses, or metrics
- **Consistent**: Easy-to-use API
- **Applicable**: Allows evaluating unintended interactions among defenses and attacks

Amulet repo: https://github.com/ssg-research/amulet

Amulet: Structure



Takeaways

Are we using the right adversary models? Needs work

Robustness against false accusations in MORs needs improvement

More generally, ML security/privacy research needs widely accepted, streamlined adversary models

Can we simultaneously deploy defenses against multiple concerns? Needs work

Important consideration but not yet sufficiently explored

Amulet: A Toolkit for exploring various attacks and defenses, available as open source

More on our ML security/privacy work at https://ssg-research.github.io/mlsec/

Takeaways

Are we using the right adversary models? Needs work

Robustness against false accusations in MORs needs improvement

More generally, ML security/privacy research needs widely accepted, streamlined adversary models

Can we simultaneously deploy defenses against multiple concerns? Needs work

Important consideration but not yet sufficiently explored

Amulet: A Toolkit for exploring various attacks and defenses, available as open source

Other research topics:
ML security/privacy:
ML ownership resolution, Conflicting ML defenses, ML property attestation, robust concept removal in gen AI
Platform security: hardware-assisted run-time security, secure outsourced computing

Open (postdoc, grad student) positions to help lead our work: ML security/privacy, platform security
https://asokan.org/asokan/research/SecureSystems-open-positions-Jan2024.php