UNIVERSITY OF WATERLOO

Aalto University

SoK: Unintended Interactions among Machine Learning Defenses and Risks

Vasisht Duddu, Sebastian Szyller, N. Asokan

Secure Systems Group

vasisht.duddu@uwaterloo.ca, contact@sebszyller.com, asokan@acm.org

Introduction

Machine Learning (ML) models are susceptible to a wide range of risks to

- Security, Privacy, and Fairness

Prior work has explored defenses to mitigate specific risks

- Defenses typically evaluated only vs. those 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?

Unintended interactions among defenses and risks

Unintended Interactions among defenses

Combining multiple defenses may result in conflicts

- Watermarking vs. adversarial training or differential privacy[1]
- ……. many other conflicts[2,3,4]

Unintended Interactions between a defense and *other* risks

An effective defense may increase or decrease susceptibility to other risks

- Limited evaluation for some risks, defenses, interactions[3,4,5] or underlying causes[3,4]
- No systematic framework to explore unintended interactions

[1] S.Szyller, N. Asokan. *Conflicting Interactions Among Protection Mechanisms for Machine Learning Models*. AAAI 2023. https://arxiv.org/abs/2207.01991
[2] Fioretto et al. *Differential Privacy and Fairness in Decision and Learning Tasks: A Survey*. IJCAI 2022. https://arxiv.org/abs/2202.08187
[3] Ferry et al. *SoK: Taming the Triangle - On the Interplays between Fairness, Interpretability and Privacy in Machine Learning.* arXiv 2024. https://arxiv.org/abs/2312.16191
[4] Gittens et al. *An Adversarial Perspective on Accuracy, Robustness, Fairness, and Privacy: Multilateral-Tradeoffs in Trustworthy ML*. IEEE Access 2024. https://ieeexplore.ieee.org/document/9933776
[5] Strobel and Shokri. *Data Privacy and Trustworthy Machine Learning*. IEEE S&P Magazine 2022. https://ieeexplore.ieee.org/document/9802763

Contributions

A systematic framework for understanding unintended interactions

- overfitting & memorization conjectured as underlying causes, exploring influencing factors

Survey of existing literature on unintended interactions

- situate existing work within our framework

Guideline to conjecture previously unexplored interactions

- empirically validation for two unexplored interactions

Background: ML risks and defenses

Defenses	Risks	
RD1 (Adversarial Training)	R1 (Evasion)	Risks to Security (Integrity)
RD2 (Outlier Removal)	R2 (Poisoning)	
RD3 (Watermarking) RD4 (Fingerprinting)	R3 (Unauthorized Ownership)	Risks to Security (Confidentiality)
PD1 (Differential Privacy)	P1 (Membership Inference) P2 (Data Reconstruction)	Risks to Privacy
?	P3 (Attribute Inference)	
?	P4 (Distribution Inference)	
FD1 (Group Fairness) FD2 (Explanations)	F (Discriminatory Behaviour)	Risks to Fairness

?: No defenses with theoretical guarantees

Overview of unintended interactions

Explore pairwise interactions between each defense and all unrelated risks:

Defenses	Risks
RD1 (Adversarial Training) RD2 (Outlier Removal)	R1 (Evasion) R2 (Poisoning)
RD3 (Watermarking) RD4 (Fingerprinting)	R3 (Unauthorized Ownership)
PD1 (Differential Privacy)	P1 (Membership Inference) P2 (Data Reconstruction) P3 (Attribute Inference) P4 (Distribution Inference)
FD1 (Group Fairness) FD2 (Explanations)	F (Discriminatory Behaviour)

Overfitting and memorization are underlying causes (conjecture)

- Effective defenses may induce, reduce or rely on overfitting or memorization
- Risks tend to exploit overfitting or memorization

Underlying causes: overfitting and memorization

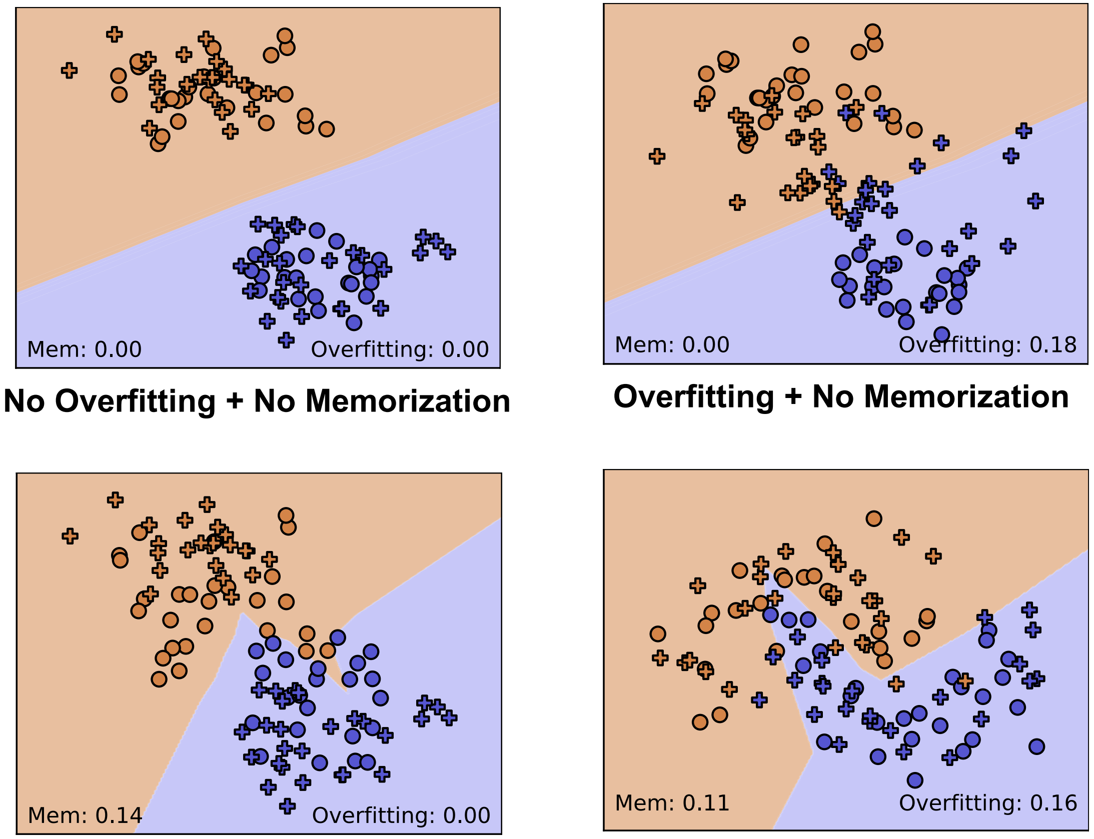
Overfitting and memorization are distinct and can occur simultaneously[1,2]

Overfitting

- Difference between train and test accuracy[3]
- Aggregate metric computed across datasets

Memorization of training data records

- Difference in model prediction on a data record with and without it in training dataset[4]
- Metric for individual data records



No Overfitting + No Memorization | **Overfitting + No Memorization**

No Overfitting + Memorization | **Overfitting + Memorization**

[1] Carlini et al. *The Secret Sharer: Evaluating and testing unintended memorization in neural networks*. USENIX Sec 2019. https://arxiv.org/abs/1802.08232
[2] Burg and Williams. *On memorization in probabilistic deep generative models.* NeurIPS 2019. https://arxiv.org/abs/2106.03216
[3] Hardt et al. *Train faster, generalize better: Stability of stochastic gradient descent.* ICML 2016. https://arxiv.org/abs/1509.01240
[4] Feldman. *Does learning require memorization? A Short Tale About a Long Tail.* STOC 2020. https://arxiv.org/abs/1906.05271

Framework: factors influencing overfitting

Bias is an error from poor hyperparameter choices for model

- High bias (smaller models) ➞ prevents learning relations between attributes and labels

Variance is an error from sensitivity to changes in the training dataset

- High variance ➞ model fits noise in training data

Tradeoffs can be balanced using:

- `D1` **Size of training data** inversely correlated with overfitting: likelihood that the model encounters a similar data record is higher
- `M1` **Model capacity** inversely correlated with overfitting if model is too simple to fit data

Framework: factors influencing memorization

D2 **Tail length of distribution** correlates with memorization of tail classes (rare or outliers)

D3 **Number of attributes** inversely correlates with memorization of individual attributes

D4 **Priority of learning stable attributes** correlates with generalization

O1 **Curvature smoothness of the objective function** results in variable memorization of data records as it determines convergence of their loss towards a minima

O2 **Distinguishability of model observables across datasets (O2.1), subgroups (O2.2), and models (O2.3)** correlates with memorization

O3 **Distance of training data to decision boundary** inversely correlates with memorization

M1 **Model capacity** Increasing capacity can increase memorization of data records

Revisiting ML risks and defenses

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]

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]

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

Dominant factors

LEGEND

O1 Curvature smoothness of the objective function
O2 Distinguishability of model observables 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 inversely
D4 Priority of learning stable attributes
M1 Model capacity

Active factors are exploited by the attacks: O1, O2, O3

Passive factors (data/model configuration): D1, D2, D3, D4, M1

Attacks often exploit dynamic factors, we deem them "dominant"

PD1 (Differential Privacy) and R1 (Evasion)→ 🔴 [1,2]

- D2 → 🟢; O1 → 🔴; O3 → 🔴

FD1 (Group Fairness) and P1 (Membership Inference) → 🔴 [3]

- D4 → 🟢; O3 → 🔴

[1] Tursynbek et al. *Robustness threats of Differential Privacy*. NeurIPS Privacy Preserving ML Workshop. 2020. https://arxiv.org/abs/2012.07828
[2] Boenisch et al.. *Gradient masking and the underestimated robustness threats of differential privacy in deep learning.* ArXiv 2021. https://arxiv.org/abs/2105.07985
[3] Chang and Shokri. *On the Privacy Risks of Algorithmic Fairness.* EuroS&P 2021. https://arxiv.org/abs/2011.03731

Group fairness (FD1) vs. data reconstruction (P2)

Conjectured Interaction from common factor:

O2.2 Distinguishability across subgroups: FD1 ↓, P2 ↑ (➞ ●)

Non-common factor: D3 # Attributes -- risk may decrease with D3

Empirical Evidence

Fair model ➞ lower attack success (confirms ●)

- Lowers distinguishability across subgroups

Metric	Baseline	Fair Model
Accuracy	84.40 ± 0.09	77.96 ± 0.58
Recon. Loss	0.85 ± 0.01	0.95 ± 0.02

Non-common factor D3

attributes = 10:

- Fair model ➞ lower attack success

attributes > 10:

- Fair model ➞ no change in attack success
  (note: # attributes do not affect accuracy drop caused by fairness)

#Attributes	Baseline		Fair Model	
	Recon. Loss	Accuracy	Recon. Loss	Accuracy
10	0.85 ± 0.01	84.40 ± 0.09	0.95 ± 0.02	78.96 ± 0.58
20	0.93 ± 0.03	84.72 ± 0.22	0.93 ± 0.00	80.32 ± 1.12
30	0.95 ± 0.02	84.41 ± 0.39	0.94 ± 0.00	79.50 ±0.91

Explanations (FD2) vs. distribution inference (P4) (1/2)

Conjectured interactions from common factor:

O2.1 Distinguishability of observables across datasets: FD2 ↑ , P4 ↑ (➞ ●)

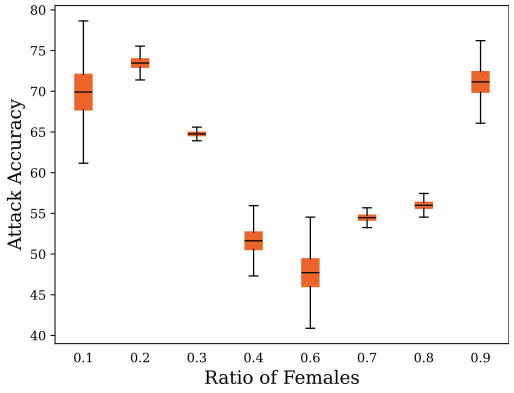
Non-common factors:

D3 **# Attributes**: risk may decrease with D3 (lower memorization)

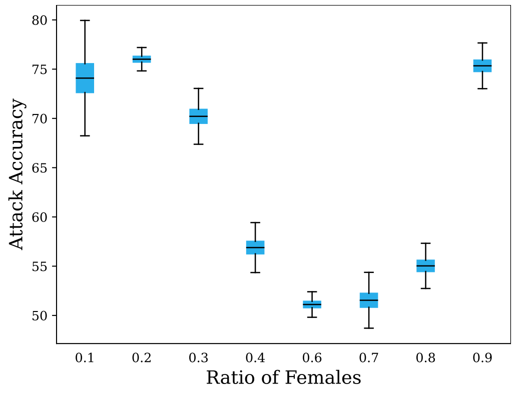
M1 **Model Capacity:** risk may increase with M1 (higher memorization)

Empirical Evidence (confirms ●)

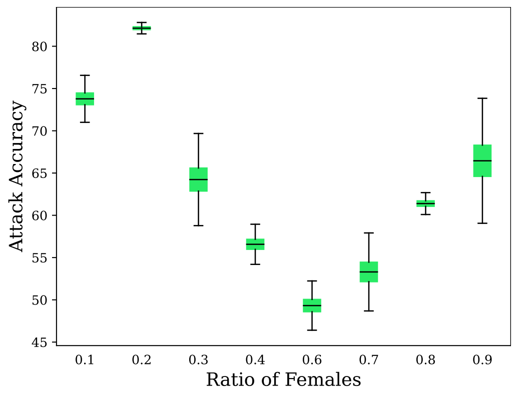
Explanations ➞ increased susceptibility to inference: attack accuracy > 50% for most ratios



Integrated Gradients



SmoothGrad



DeepLift

Explanations (FD2) vs. distribution inference (P4) (2/2)

Non-common factor `D3` (# Attributes): More attributes ➞ lower attack success

# Attributes	Integrated Gradients	DeepLift	SmoothGrad
15	81.07 ± 2.13	78.74 ± 1.66	65.40 ± 1.39
25	66.09 ± 0.95	73.64 ± 1.38	59.42 ± 1.09
35	50.43 ± 0.59	59.93 ± 2.81	56.78 ± 1.93

Non-common factor `M1` (Model Capacity): Higher capacity ➞ higher attack success

# Parameters	Integrated Gradients	DeepLift	SmoothGrad
5.7K	47.57 ± 4.25	49.19 ± 2.75	53.26 ± 0.10
44K	53.29 ± 3.65	50.86 ± 3.24	62.40 ± 0.95
274K	62.60 ± 2.74	67.73 ± 1.69	70.21 ± 0.73
733K	69.90 ± 3.24	73.78 ± 1.03	74.09 ± 2.17

Exceptions to guideline

Differences in adversary models can change the interaction type

- **RD1 (Adversarial training) and R3 (Unauthorized Model Ownership)**
  - Guideline predicts ➞ ● (M1 but not dominant)
  - If adversary is malicious suspect➞ ●[1]; If adversary is malicious accuser➞ ●[2]
- **PD1 (Differential privacy) and P4 (Distribution Inference)**
  - Guideline predicts ➞ ● (O2.1) which matches with empirical evidence[3]
  - If adversary knows victim is DP-trained, they can DP-train shadow models➞ ●[3]
- **FD1 (Group fairness) and P3 (Attribute Inference)**
  - Guideline predicts ➞ ● (O2.2) which matches with empirical evidence[4]
  - If adversary knows fairness algorithm, they can calibrate their attack➞ ●[5]

Some defenses and risks have too few factors

- RD2 (Outlier removal), R2 (Poisoning), R3 (Unauthorized model ownership)

[1] Khaled et al. *Careful What You Wish For: On the Extraction of Adversarially Trained Models.* PST 2022. https://arxiv.org/abs/2207.10561
[2] Liu et al. False Claims against Model Ownership Resolution. Usenix SEC 2024. https://arxiv.org/abs/2304.06607
[3] Suri et al. *Dissecting Distribution Inference*. SatML 2023. https://arxiv.org/abs/2212.07591
[4] Aalmoes et al. *On the alignment of Group Fairness with Attribute Privacy.* ArXiv 2022. https://arxiv.org/html/2211.10209v2
[5] Ferry et al. *Exploiting Fairness to Enhance Sensitive Attributes Reconstruction.* SatML 2023. https://arxiv.org/abs/2209.01215

Current work

Unexplored Interactions:

- RD1 (Adversarial Training) ➞ P3 (Attribute Inference)
- RD2 (Outlier Removal) ➞ R3 (Unauthorized Model Ownership)
- RD2 (Outlier Removal) ➞ P2 (Data Reconstruction)
- RD2 (Outlier Removal) ➞ P4 (Distribution Inference)
- RD3 (Watermarking) ➞ R1 (Evasion)
- RD4 (Fingerprinting) ➞ R2 (Poisoning)
- RD4 (Fingerprinting) ➞ P2 (Data Reconstruction)
- RD4 (Fingerprinting) ➞ P3 (Attribute Inference)
- RD4 (Fingerprinting) ➞ P4 (Distribution Inference)
- PD1 (Differential Privacy) ➞ R3 (Unauthorized Model Ownership)
- FD1 (Group Fairness) ➞ R3 (Unauthorized Model Ownership)

Developing a software framework for systematic empirical evaluation

Need to understand impact of defense/risk variants on their interactions

Takeaways

Unintended interactions are an important concern in practice

Common influencing factors can help identify such interactions

ML Sec/Priv Research @ Secure Systems Group
https://ssg-research.github.io/mlsec/

Current: systematic empirical evaluation of unintended interactions

Future: how to design defenses to minimize increases in *other* risks?

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