Secure Systems Group, University of Waterloo[+], Aalto University[†]

Vasisht Duddu[+], Sebastian Szyller[†], N. Asokan[+,†]

SoK: Unintended Interactions among Machine Learning Defenses and Risks

Motivation

- ML models susceptible to different risks to security, privacy, and fairness
- Defenses designed against specific risks

But may also impact unrelated risks
Unintended interactions

- No systematic framework to understand them

Unintended interactions

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)

Conjectured causes: overfitting, memorization

Framework: Underlying causes

Overfitting: Difference in train and test accuracy

Factors: Trainset size (D1); Model capacity (M1)

Memorization: Difference in model prediction on data record w/ and w/o it in training dataset

Influencing factors:

- **Dataset**: Tail length of distribution (D2); number of attributes (D3); priority of learning stable attributes (D4)
- **Objective function:** curvature smoothness (O1); distinguishability of observables across datasets (O2.1), subgroups (O2.2), models (O2.3); distance to decision boundary (O3)
- **Model:** same as M1

Situating prior work in framework

- Risk increases→●, decreases→●
- Interaction unexplored→●
- Factors evaluated: empirical→●, theoretical→⊙, conjectured→○

Defenses	Risks	I	OVFT D1	Memorization D2	D3	D4	O1	O2	Both O3	M1	References
RD1 (Adversarial Training)	R1 (Evasion)	●		●			●		●	●	[96], [108], [182], [202]
	R2 (Poisoning)	●									[161], [179]
	R3 (Unauthorized Model Ownership)	●	○								[90] ([101]; ●)
	P1 (Membership Inference)	●	⊙, ●					1: ●		●	[71], [152]
	P2 (Data Reconstruction)	●				○				●	[117], [204]
	P3 (Attribute Inference)	●									
	P4 (Distribution Inference)	●				○					[156]
	F (Discriminatory Behaviour)	●		⊙, ●							[16], [38], [75], [105]
RD2 (Outlier Removal)	R1 (Evasion)	●									[63]
	R2 (Poisoning)	●									[163]
	R3 (Unauthorized Model Ownership)	●									
	P1 (Membership Inference)	●		●							[26], [47]
	P2 (Data Reconstruction)	●									
	P3 (Attribute Inference)	●		●							[82]
	P4 (Distribution Inference)	●									
	F (Discriminatory Behaviour)	●	●	○							[142]

Guideline for conjectures

Defences (<↑ or ↓>, <f>)

RD1 (Adversarial Training):
- D1 ↑, $|\mathcal{D}_{tr}|$ [170]
- D2 ↓, tail length [16], [75]
- D4 ↑, priority for learning stable attributes [170]
- O1 ↑, curvature smoothness [108]
- O2.1 ↑, distinguishability in data records inside and outside $\mathcal{D}_{tr}$ [152]
- O3 ↑, distance to boundary for most $\mathcal{D}_{tr}$ data records [185]
- M1 ↑, model capacity [108]

RD2 (Outlier Removal):
- D2 ↑, tail length [175]

RD3 (Watermarking):
- D2 ↑, tail length [102]
- O2.3 ↓, distinguishability in observables between watermarks between f_θ and f_θ^{der}, but distinct from independent models
- M1 ↑, model capacity [3]

Risks (<↑ or ↓>, <f>)

R1 (Evasion):
- D2 ↑, tail length [96], [182]
- O1 ↓, curvature smoothness [108]
- O3 ↓, distance of $\mathcal{D}_{tr}$ data records to boundary [171]

R2 (Poisoning):
- D2 ↑, tail length [17], [102], [127]
- M1 ↑, model capacity [3]

R3 (Unauthorized Model Ownership):
- M1 ↓, model capacity [93], [124]

P1 (Membership Inference):
- D1 ↓, $|\mathcal{D}_{tr}|$ [144], [193]
- D2 ↑, tail length [25], [26]
- D4 ↓, priority for learning stable attributes [109], [164]
- O2.1 ↑, distinguishability for data records inside and outside $\mathcal{D}_{tr}$ [144]
- O3 ↓, distance to decision boundary [145]
- M1 ↑, model capacity [48], [152]

P2 (Data Reconstruction):

Effectiveness of defense correlates with factor

Change in factor (<f>) correlates with risk

- ↑: positive correlation; ↓: negative correlation

Use arrows for <defense, f> and <f, risk>:

- If (↑,↑) or (↓,↓) → ●; else (↑,↓) or (↓,↑) → ●

Conjecture is:

- unanimous if all factors agree, or
- determined by dominant factor (O1, O2, O3)

Non-common factors may affect interaction

Aalto University

ssg-research.github.io/mlsec/interactions

UNIVERSITY OF WATERLOO