UNIVERSITY OF WATERLOO | DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

CrySP

Unintended Interactions in Protecting ML Models

Vasisht Duddu (Joint work with Sebastian Szyller, Rui Zhang, N. Asokan)

During deployment, simply designing defenses against individual ML risks is not sufficient

- **R1: Defense against one risk may increase/decrease susceptibility to other unrelated risks**
- **R2: Conflicts among defenses being combined against multiple risks, degrades effectiveness**

Overarching concerns during model deployment ➞ "Meta-Concerns"

R1: Defenses vs. Unrelated Risks[1]

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 and Memorization

Defense ➞ overfitting and memorization ➞ Risks

Framework: Influencing Factors

Factors influencing overfitting and memorization likely influence interactions among defenses and risks

Size of training dataset
Tail Length of Distribution
Priority of Learning Stable Attributes
Number of Input Attributes
Curvature Smoothness
Distinguishability in model observables
Distance to decision boundary
Model Capacity

Guideline to Predict Interactions

Defenses	Risks
RD1 (Adversarial Training) • ↑, Size of training data • ↓, Tail length of distribution •	**R1 (Evasion)** • ↑, Tail length of distribution • ↓, Curvature Smoothness • ...
RD2 (Outlier Removal) •	**R2 (Poisoning)** • ....

Check how:

- **Defense effectiveness correlates with factor**
- **Change in factor correlates with risk susceptibility**
- ↑: positive correlation; ↓: negative correlation

⇒ Infer how defense effectiveness correlates with risk
(↑,↑) or (↓,↓) ➞ 🔴 and (↑,↓) or (↓,↑) ➞ 🟢

Evaluation and Results

- **Identified two unexplored interactions**
- **Predicted interactions using guideline**
- **Validated prediction from guideline empirically**

R2: Conflicts among ML Defenses[2]

Protect against multiple risks by combining defenses

Effective combination ➞ No drop in defense effectiveness

Need principled combination technique to identify if combination is effective (⇒ no conflict)

Requirements

- **Accurate**: Correctly identify if combination is effective
- **Scalable**: Allows combining more than two defenses
- **Non-invasive**: No changes to defenses being combined
- **General**: Applicable to different types of defenses

Existing Techniques and Limitations

Optimization: Not scalable, not general, invasive

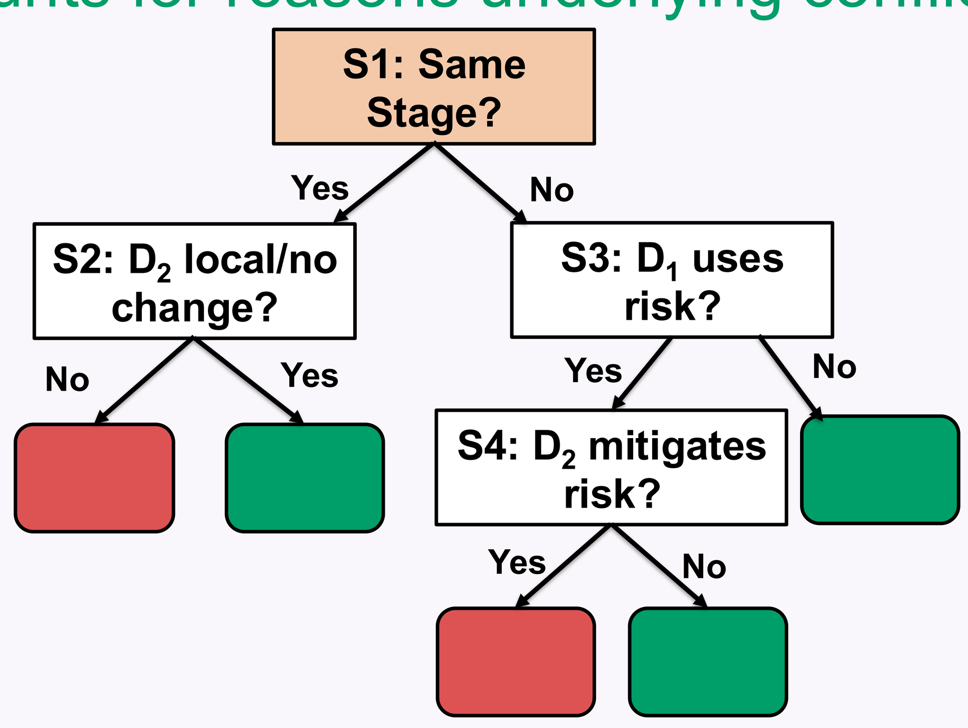
Mutually Exclusive Placement (Naïve): Not accurate

- Changes by second defense overrides first defense
- Second defense minimizes risk used by first defense

Def\Con

Improves accuracy of naïve technique

- Check position of defenses, and if mechanisms interfere
- Explicitly accounts for reasons underlying conflicts



Evaluation and Results

Explored combinations (ground truth from prior work)

Def\Con: 90% (7/8) vs. Naïve: 40% (4/8)

Unexplored combinations (ground truth from evaluation)

Def\Con: 81% (27/30) vs. Naïve: 36% (18/30)

Def\Con is more accurate than naïve technique, scalable, non-invasive, and general

[1] *SoK: Unintended Interactions among Machine Learning Defenses and Risks*. IEEE S&P 2024. (Distinguished Paper Award) 🏆

[2] *Combining Machine Learning Defenses without Conflicts*. TMLR 2025.