Trustworthy Deployment of Machine Learning Systems

Vasisht Duddu vasisht.duddu@uwaterloo.ca vasishtduddu.github.io

About me

Ph.D. student, University of Waterloo (Canada)

Advisor: N. Asokan

Previously: *Masters @ University of Waterloo, Undergraduate @ IIIT-Delhi, India*

Security and privacy researcher working on making ML systems trustworthy

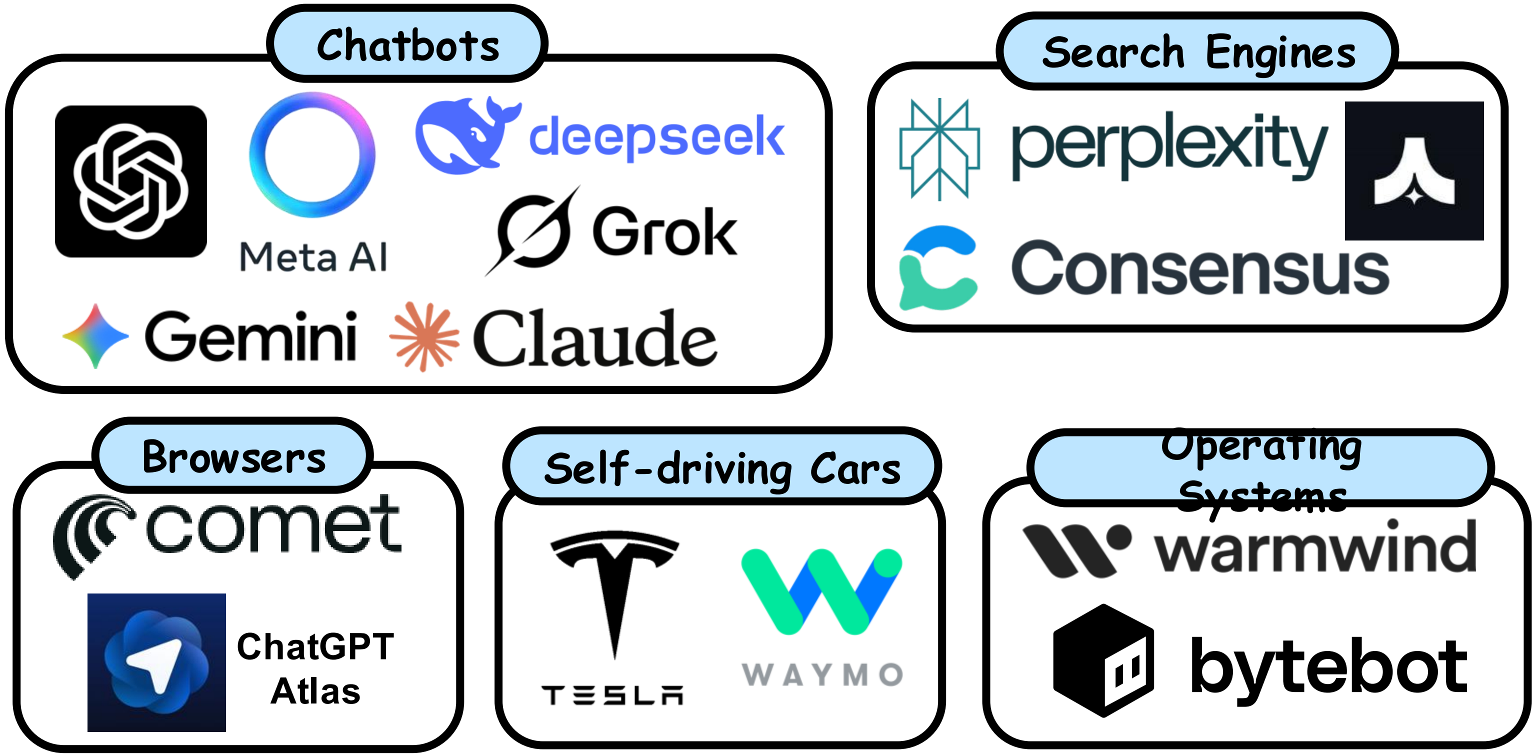
- IBM Ph.D. Fellowship (2024)
- Distinguished Paper @ IEEE S&P (2024)
- Best Paper @ ACM CODASPY (2025)
- Technology transfer to Intel (2025)

https://vasishtduddu.github.io/

Introduction

Significant utility improvement in machine learning (ML) → Wide-scale deployment

- Client-facing services (e.g., chatbots, search engines, browsers)
- High-stakes applications (e.g., healthcare, criminal justice)
- Part of larger systems (e.g., operating systems, autonomous vehicles)



Deployment Concerns

Infrastructure
Latency, throughput, interoperability, scalability,….

Model Design
Utility, generalization, hyperparameter tuning, data processing

Environment
Carbon emissions, power consumption, water usage

Safety Risks
Misinformation, surveillance, misalignment, cyberattacks, ….

Adversarial Risks
Security, privacy, fairness, transparency, unintended interactions

Governance
Accountability, regulatory compliance, verifiability

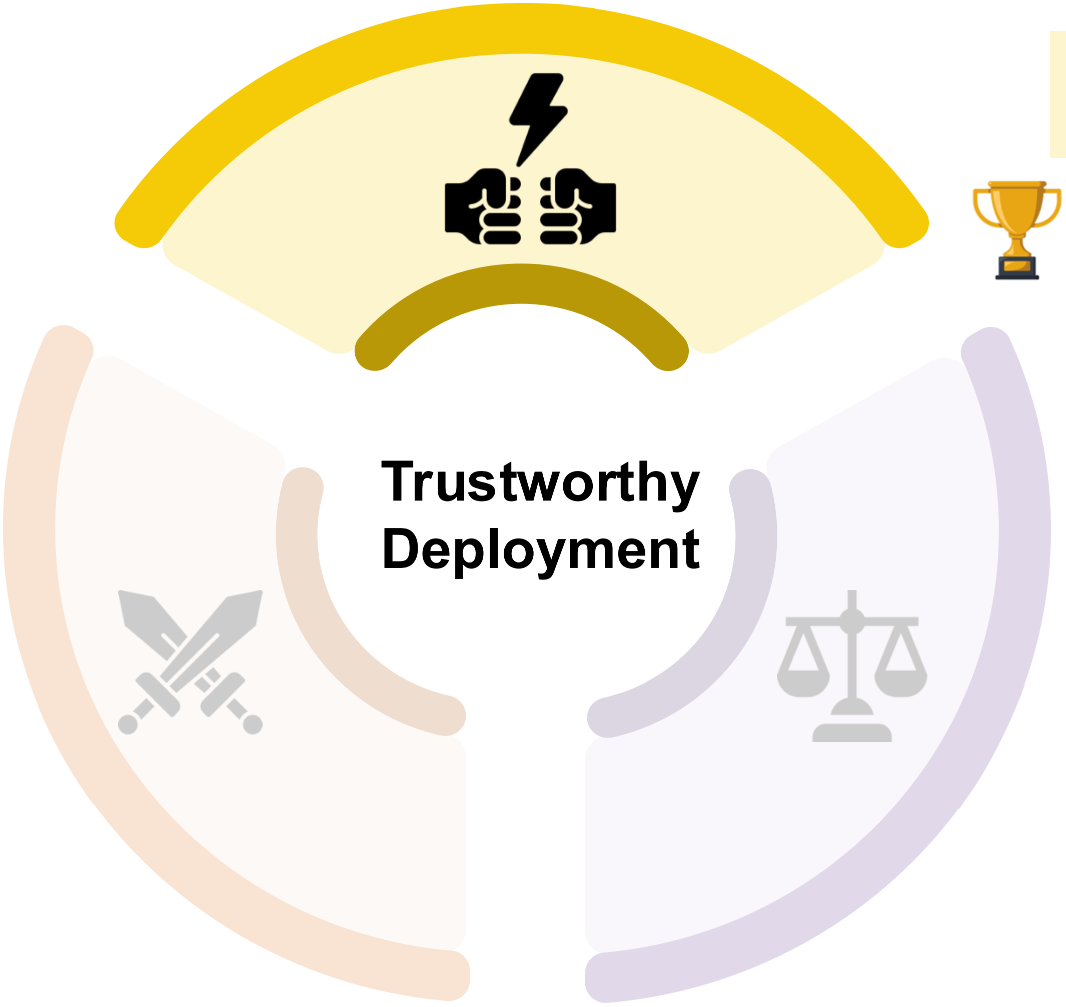
Trustworthy Deployment

Talk Overview

Exploring “Meta-Concerns”

CIKM’22, WISE’24, S&P’24b, TMLR’25, ArXiv’25a

Distinguished Paper @ IEEE S&P’24
Technology Transfer to Intel



Identifying and Mitigating Risks

Security, Privacy, Fairness, Transparency

MobiQuitous’20, SAC’22, S&P’24a, CODASPY’25, ICML’25, PETS’26

Best Paper @ ACM CODASPY’25
Oral @ AAAI PPAI Workshop’25

Enabling Governance

CCS’23, ESORICS’24, CODASPY’25, ArXiv’25b

Overview of ML Risks and Defenses

Evasion and Jailbreak
Perturb inputs to force misclassification or forbidden output

➜ Adversarial training and robust alignment

Unauthorized Model Ownership
Steal functionality of target model

➜ Watermarking and fingerprinting

Poisoning/Backdoor
Manipulate training data or model or training to degrade utility or generate adversary-chosen output

➜ Outlier robustness (data sanitization, finetuning, pruning)

Unauthorized Data Usage
Use of copyrighted or personal data without consent

➜ Watermarking

Security

Inference Attacks
Infer sensitive information from model: membership, attribute, distribution inference, data reconstruction

➜ Differential privacy

Privacy

Bias and Incomprehensibility
Model behaves differently across demographic subgroups, and unclear why model made specific predictions

➜ Individual and group fairness; Post-hoc explanations

Fairness

Exploring "Meta-Concerns": Contributions

Not enough to design effective defenses against **individual risks**

Practitioners need to protect against multiple risks simultaneously

Problem 1

Unintended Interactions among Defenses and Risks

Why does defense increase or decrease unrelated risks?

CIKM'22 WISE'24

S&P'24

Distinguished Paper

Problem 2

Conflicts among Defenses when Combined

How can defenses be combined without conflicts?

TMLR'25

Problem 3

Colluding Adversaries in ML Pipelines

How can adversaries collude by exploiting one risk to increase others?

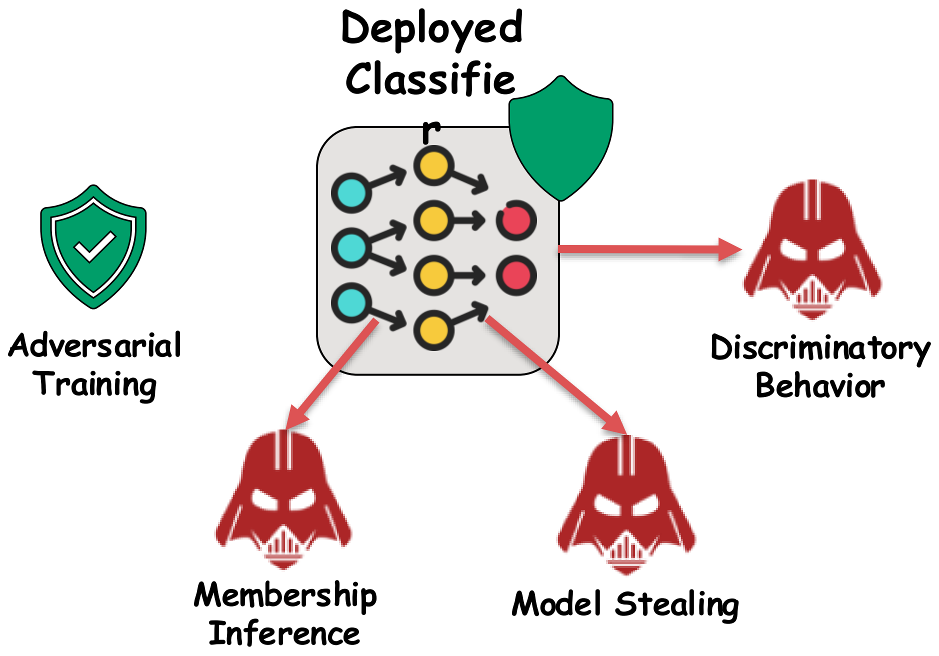
ArXiv'26

(Under submission)

Guideline for practitioners to predict unintended interactions or conflicts without expensive evaluation

Problem 1: Defenses vs. Unrelated Risks

S&P'24 Distinguished Paper



Example

Adversarial training increases membership inference, model stealing, and discriminatory behavior[2,3]

Prior work limited to specific risks and defenses[1,2,3]
No systematic framework to study underlying reasons

Conjecture: Overfitting and memorization are underlying causes

- Effective defenses influence overfitting or memorization
- Risks tend to exploit factors influencing overfitting or memorization

[1] Ferry et al. *SoK: Taming the Triangle - On the Interplays between Fairness, Interpretability and Privacy in Machine Learning*. ArXiv. 2024.
[2] Gittens et al. *An Adversarial Perspective on Accuracy, Robustness, Fairness, and Privacy: Multilateral-Tradeoffs in Trustworthy ML*. IEEE Access. 2024.
[3] Strobel and Shokri. *Data Privacy and Trustworthy Machine Learning*. IEEE S&P Magazine. 2022.

Factors Influencing Overfitting and Memorization

Curvature smoothness of the objective function

Distinguishability across (a) datasets, (b) subgroups, and (c) models

Distance of training data to decision boundary

(Objective function-related)

Size of training data

Tail length of distribution

Number of attributes

Priority of learning stable attributes

(Dataset-related)

Model capacity

(Model-related)

Guideline to Predict Unintended Interactions

Effectiveness of defense correlates with change in factor

Change in factor correlates with change in susceptibility to risk

- Identify correlations with factors for all defenses and risks
- **Example**: Group Fairness vs. Data Reconstruction

Conjecture → Group fairness reduces data reconstruction

Condition → Conjecture holds for less attributes

Group Fairness (Defense)

Experiment Setup
Train neural network on CENSUS (tabular data) for binary classification of income > \$50K

Recon. Loss: L_2(input, recon. input) [lower better]
Fairness: p%-rule > 80% (demographic parity)

↓ (Number of input attributes)

↑ (Distinguishability of outputs across subgroups)

# Input Attributes	Baseline	Fair Model
	Recon. Loss	Recon. Loss
10	0.85 ± 0.01	0.95 ± 0.02
20	0.93 ± 0.03	0.93 ± 0.00
30	0.95 ± 0.02	0.94 ± 0.00

For common factor, do arrows (↓,↓)?

No

Attack less effective with fairness

with defense

Attack ineffective for # attributes >10

factors

Positive correlation (↑); Negative correlation (↓)

Validating Guideline

Apply guideline to two unexplored interactions and empirically validate them

- **Example 1:** Group fairness decreases data reconstruction
- **Example 2:** Model explanations leaks distributional properties of training data

Validate guideline by comparing conjectures with prior work

Exceptions

- Differenc … n
- Some de … d

Takeaway

Unintended interactions are important for practical deployment and practitioners can study them using underlying factors

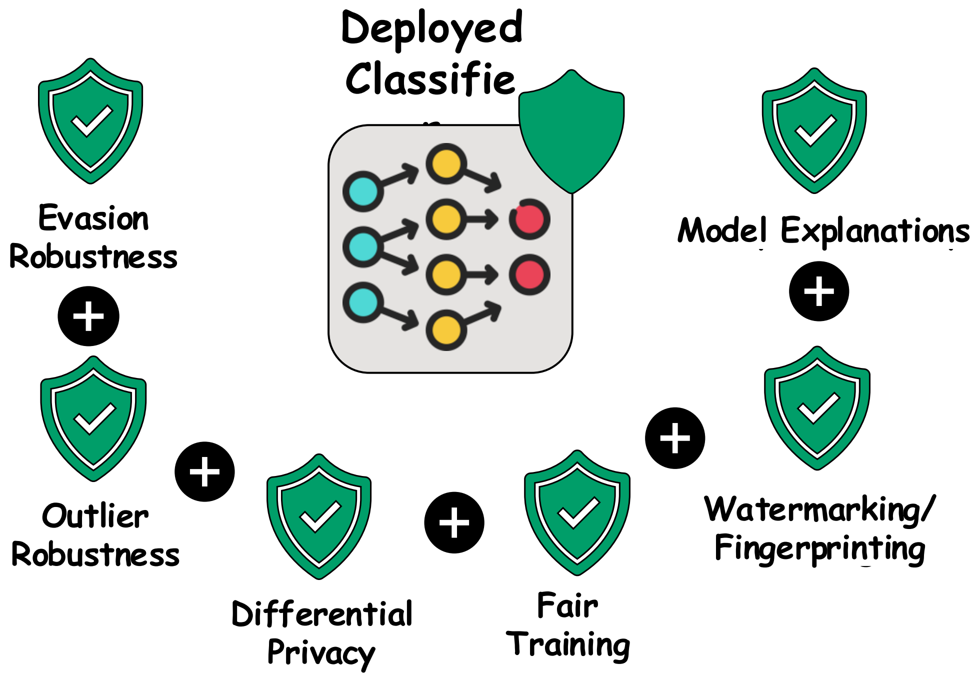
First step towards understanding interactions and further work required

Recourse for Practitioners

Tune individual factors for specific interactions to reduce unintended increase in risks

Problem 2: Protection Against Multiple Risks

TMLR'25



Effectively combine defenses to protect against multiple risks

- Defense effectiveness before and after combination is same
- **Problem** → Conflicting objectives among defenses[1,2,3,4]

Need principled combination technique

- Modify existing defenses to combine effectively
- Identify if defenses can be combined without conflict

[1] Szyller and Asokan. *Conflicting Interactions Among Protection Mechanisms for Machine Learning Models*. AAAI. 2023.
[2] Fioretto et al. *Differential Privacy and Fairness in Decision and Learning Tasks: A Survey*. IJCAI. 2022.
[3] Ferry et al. *SoK: Taming the Triangle - On the Interplays between Fairness, Interpretability and Privacy in Machine Learning*. ArXiv. 2024.
[4] Gittens et al. *An Adversarial Perspective on Accuracy, Robustness, Fairness, and Privacy*. IEEE Access. 2024.

Desiderata: Ideal Combination Technique

Accurate	Correctly identifies whether combination is effective or not
Scalable	Allows combining more than two defenses
Non-Invasive	Requires no changes to defenses being combing
General	Applicable to different types of defenses

Limitations of Prior Work

Optimization Techniques[1,2]

Game-theory, regularization, constraint solving, …

Accurate | Not Scalable | Invasive | Not General

Trade-off between effectiveness and utility

Optimization specific to combinations

Mutually Exclusive Placement[3,4]

(aka naïve technique)

Defenses in different stages are non-conflicting

Not Accurate | Scalable | Non-invasive | General

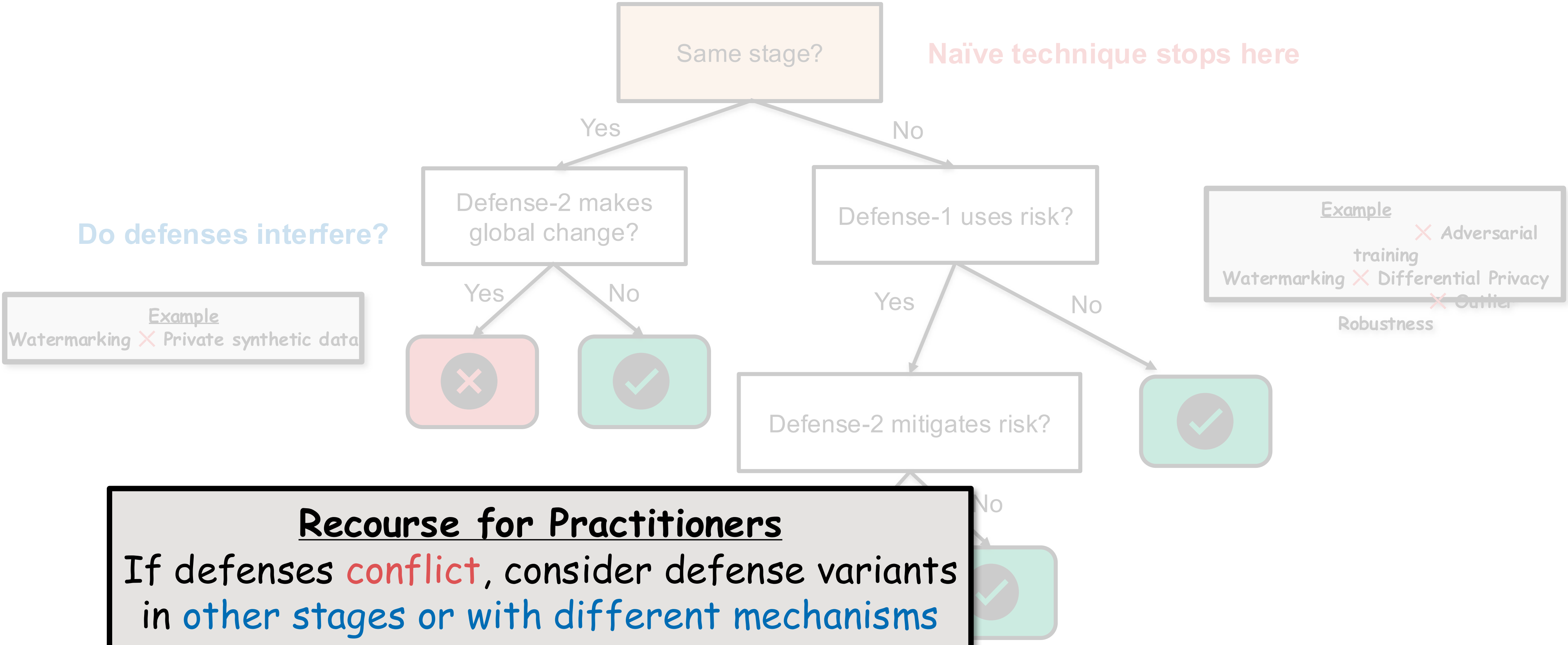
Incorrect non-conflicting same-stage and conflicting different-stage defenses

Naïve technique is promising but not accurate

Can we improve accuracy by accounting for reasons underlying conflicts?

[1] Wu et al. *Augment then smooth: Reconciling differential privacy with certified robustness*. TMLR. 2024.
[2] Tran et al. Differentially private and fair deep learning: A Lagrangian dual approach. AAAI. 2021.
[3] Szyller and Asokan. *Conflicting Interactions Among Protection Mechanisms for Machine Learning Models*. AAAI. 2023.
[4] Yaghini et al. *Learning with Impartiality to Walk on the Pareto Frontier of Fairness, Privacy and Utility*. ArXiv. 2023.

Def\Con: Design



Recourse for Practitioners

If defenses conflict, consider defense variants in other stages or with different mechanisms

Def\Con: Evaluation

Accuracy Identify defense variants in different stages → 38 pairwise combinations

Eight combinations as ground truth from prior work

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

Used empirical ev... truth

- Def\Con: 81% (2...

Takeaway

Existing defenses can be effectively combined by predicting whether defenses conflict

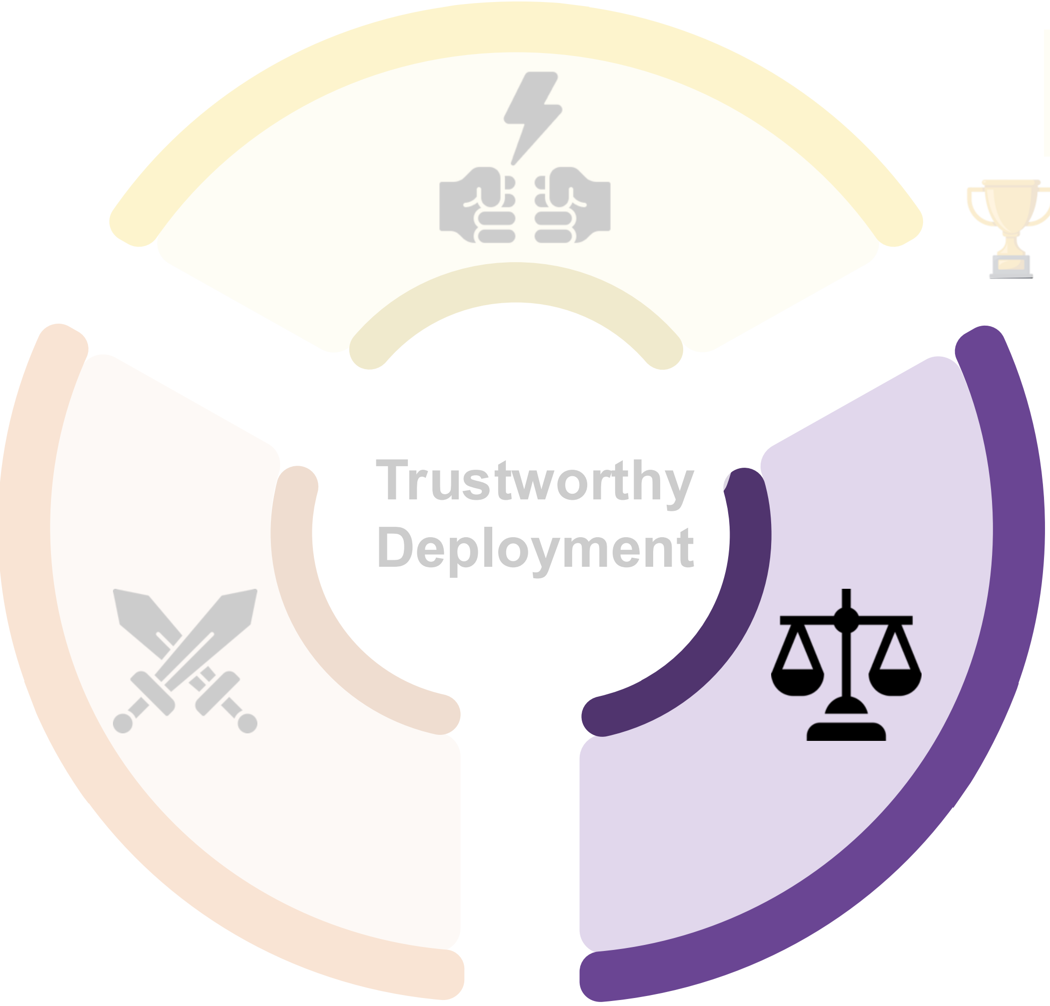
Scalable	Non-invasive	General
Can combine more than two defenses	Not modifying existing defenses	DEF\CON independent of defenses

Talk Overview

Exploring “Meta-Concerns”

CIKM’22, WISE’24, S&P’24b, TMLR’25, ArXiv’25a

Distinguished Paper @ IEEE S&P’24
Technology Transfer to Intel



Identifying and Mitigating Risks

Security, Privacy, Fairness, Transparency

MobiQuitous’20, SAC’22, S&P’24a, CODASPY’25, ICML’25, PETS’26

Best Paper @ ACM CODASPY’25
Oral @ AAAI PPAI Workshop’25

Enabling Governance

CCS’23, ESORICS’24, CODASPY’25, ArXiv’25b

Enabling Governance: Contributions

Technical Mechanisms to Ensure Accountability

How can we design mechanisms to attest ML properties?

ESORICS'24 CODASPY'25

Human-centered Studies to Inform Practitioners

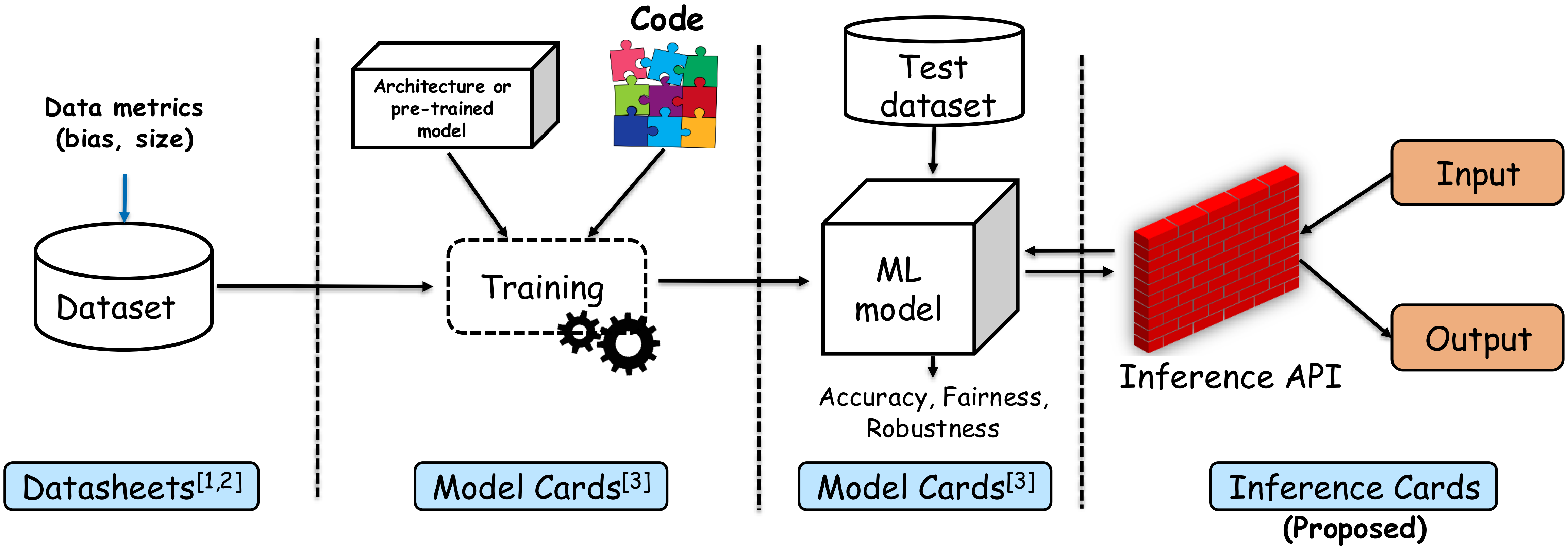
Can user expectations and perceptions inform defenses and deployment?

CCS'23 ArXiv'26

(Under submission)

Advertising ML Properties for Transparency

ESORICS'24
CODASPY'25



Collectively, refer to them as "ML property cards"

[1] Gebru et al. *Datasheets for datasets*. Communications of ACM. 2021.
[2] Pushkarna et al. *Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI*. FaccT. 2022.
[3] Mitchell et al. *Model Cards for Model Reporting*. FaccT. 2019.

Need Verifiable ML Property Cards

Malicious prover can make false claims about model or data (e.g., HuggingFace[1])

Prover (model trainer/owner) needs to convince Verifier about:

- Correct execution of ML operations (accountability)

ML property attestation[2]

- Prover (e.g., model trainer) demonstrates properties to Verifier (e.g., regulator, customer)
- Without revealing proprietary model and training data → Confidentiality

[1] Mithril-Security. *PoisonGPT: How to poison LLM supply chain on HuggingFace*. 2023.
[2] Duddu et al. *Attesting Distributional Properties of Machine Learning Training Data*. ESORICS. 2024.

Desiderata: ML Property Attestation Mechanism

Effective	Correctly estimate ML properties
Efficient	Incur low computation overhead compared to ML operations
Versatile	Support various ML properties for training, evaluation, inference
Scalable	Attestations can be efficiently checked by multiple verifiers
Robust	Resist evasion of attestations by malicious provers

Limitations of Software-based Attestations

ML-based Attestations

Examples: Proof of learning[1],
Re-purposing privacy attacks[2]

Statistical techniques and ML models for auditing

Not Effective[2] | Efficient | Versatile | Scalable | Not Robust[2,3,4]

Cryptographic Attestations

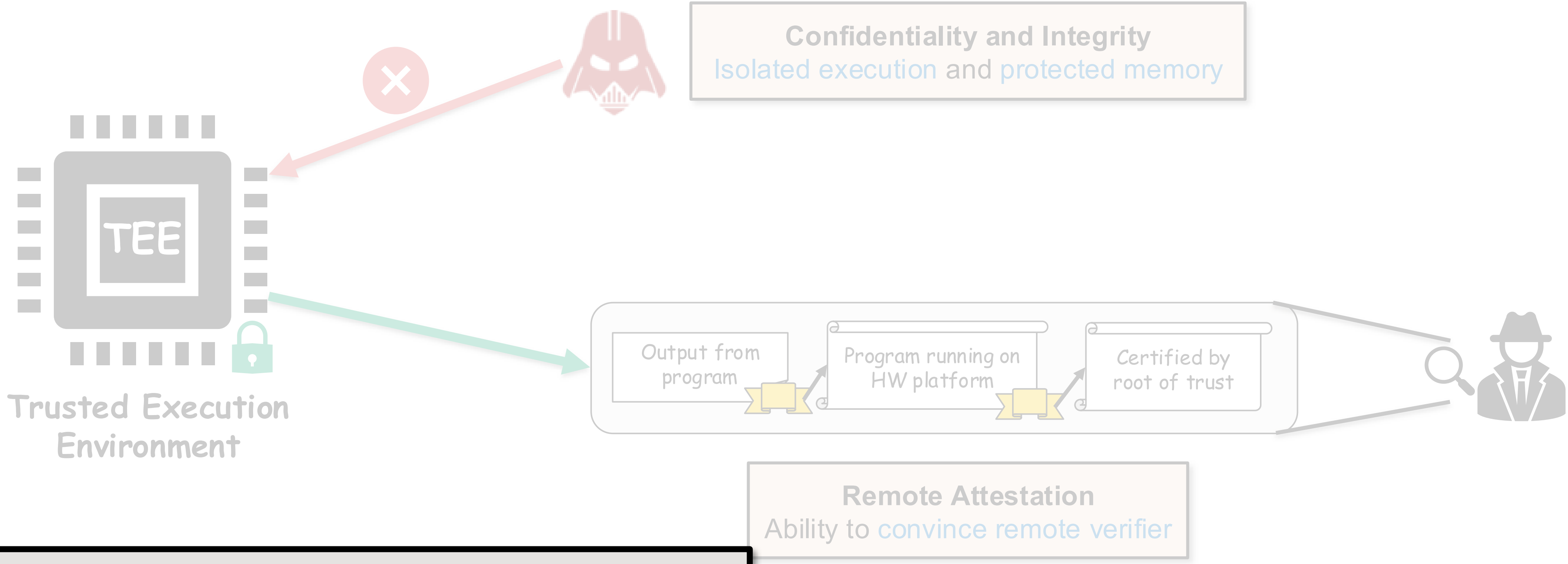
Examples: Multi-party computation[2],
Zero-knowledge proofs[5,6]

Design protocols using cryptographic primitives

Effective | Inefficient[2] | Not Versatile | Scalable | Robust

[1] Jia et al. *Proof of Learning: Definitions and Practice*. IEEE S&P. 2021.
[2] Duddu et al. *Attesting Distributional Properties of Machine Learning Training Data*. ESORICS. 2024.
[3] Zhang et al. *"Adversarial Examples" for Proof- of-Learning*. IEEE S&P. 2022.
[4] Fang et al. *Proof of Learning is more Broken than You Think*. IEEE EuroS&P. 2023.
[5] Sun et al. *zkLLMs: Zero Knowledge Proofs for Large Language Models*. ACM CCS. 2024.
[6] Abbaszadeh et al. *Zero-Knowledge Proofs of Training for Deep Neural networks*. ACM CCS. 2024.

Hardware-assisted Attestations



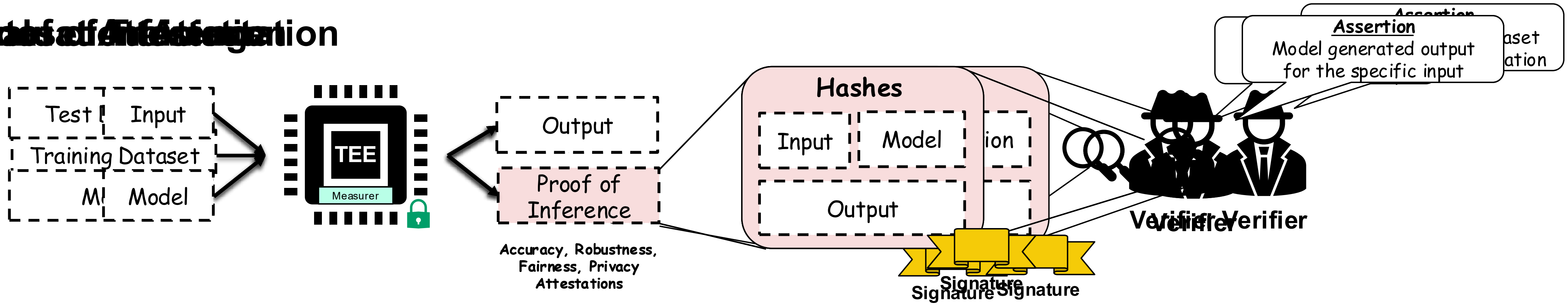
Can we adapt remote attestation to efficiently[1,2] demonstrate ML properties?

[1] Google Cloud Team. We tested Intel's AMX CPU accelerator for AI and here's what we learned. 2024.
[2] Zhu et al. Confidential Computing on Nvidia's H100 GPU: A Performance Benchmark Study. ArXiv. 2024.

Laminator Framework

Use TEEs to furnish ML property attestations

- Measurer script within TEE measures desired property



Laminator: Evaluation

Efficiency — Laminator incurs low overhead for attestations (<2%)

Effective	Scalable	Versatile	Robust
Measurer script correctly	Attestations can be checked	Any property specified in	Inherited from TEE's

Takeaway

Hardware-assisted TEEs are promising to effectively and efficiently furnish attestations and enable accountability in ML pipelines

Laminator meets all requirements ... furnish verifiable ML property cards

Future Work: Trustworthy and Verifiable AI Agents

Identifying and Mitigating Risks

- Systematic evaluation of emerging risks (e.g., alignment faking)
- Revisiting systems and network security risks and principles in AI ecosystem

Mitigating Risks

- Applying contextual integrity to evaluate privacy
- Control unintended behaviors using interpretability and model editing

Meta-Concerns

- Robust alignment with human expectations despite conflicts
- Emergent misalignment (fine-tuning on narrow task → misalignment)

Enabling Governance

Extending attestations for AI ecosystem

- (Runtime) Attestations for agents
- Attestations for properties of ecosystem
- Formal verification of ecosystem components

Summary

“Meta-concerns” are important in practice while protecting against multiple risks

- Defense may increase or decrease susceptibility to other risks
- Avoiding conflicts while combining defenses

Hardware-assisted TEEs are useful for attesting ML operations

---

My other research on identifying and mitigating risks (not covered):

- First work to identify privacy risks in graph-based models **MobiQuitous’20** **>180 citations**
- First fingerprinting scheme for graph-based models **S&P’24**
- Robust suppression of inappropriate/unauthorized outputs **ArXiv’25** **CODASPY’25** **Best Paper**
- Contextual integrity for language models **ICML’25** **PETS’26**
- Mechanistic interpretability to reduce PII leakage **EACL Findings’26**