Panel Discussion

Trustworthy Machine Learning: Challenges and Opportunities

Cho-Jui Hsieh
UCLA

Pin-Yu Chen
IBM Research

Soheil Feizi
University of Maryland, College Park

Sijia Liu
Michigan State University

Machine Learning Summer School 2021, Taipei

Cho-Jui Hsieh

Short bio

- B.S and M.S in Computer Science, National Taiwan University
- Ph.D. in Computer Science, University of Texas at Austin
- Assistant Professor, University of California at Davis, 2015.8 -- 2018.10
- Assistant Professor, UCLA, 2018.11 -- now
- Visiting Scholar:
    - Google: 2018 -- 2020
    - Amazon A9: 2020 -- 2021

Adversarial Robustness of ML Models

natural example

adversarial example

Prediction:
Bagel

Imperceptible Adversarial Perturbation

Prediction:
Grand Piano

Original Top-3 inferred captions:

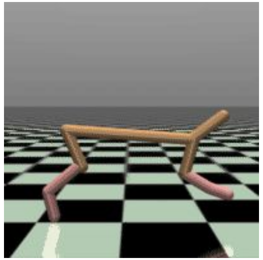
1. A close up of a giraffe with trees in the background
2. A close up of a giraffe near a fence
3. A close up of a giraffe near a tree

Adversarial Keywords:
“soccer”, “group” and “playing”

Adversarial Top-3 captions:
(targeted keyword method)

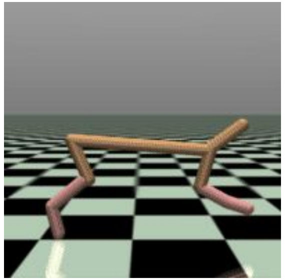
1. A group of young men playing a game of soccer.
2. A group of people playing a game of soccer.
3. A group of people playing a game of baseball.

Robustness in Reinforcement Learning (NeurIPS ‘20, ICLR ‘21)



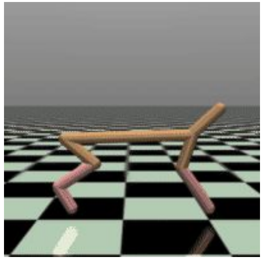
Reward:
+7094

Normal agent



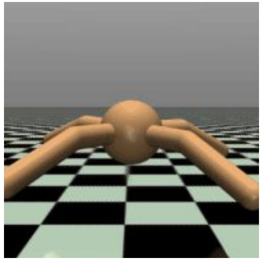
-743

Normal agent under **Optimal** attack



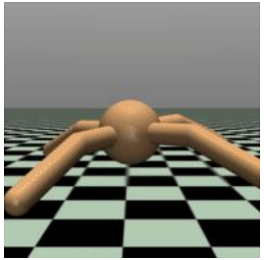
+5250

ATLA agent under optimal attack



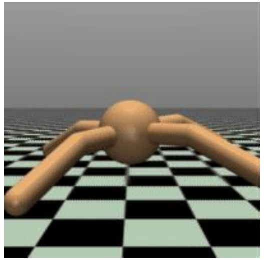
+5358

Normal agent



-1141

Normal agent under **Optimal** attack
(move opposite to the goal)



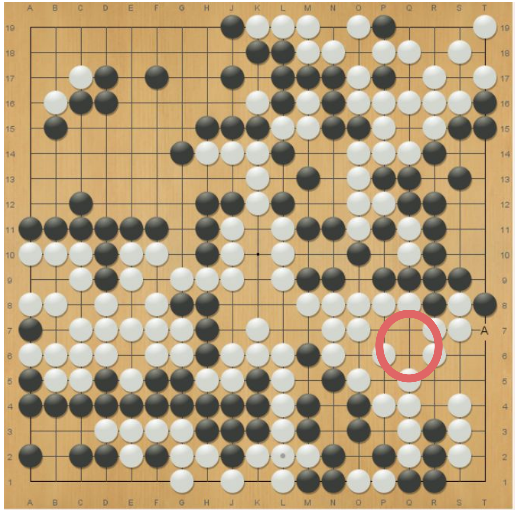
+3835

ATLA agent under optimal attack

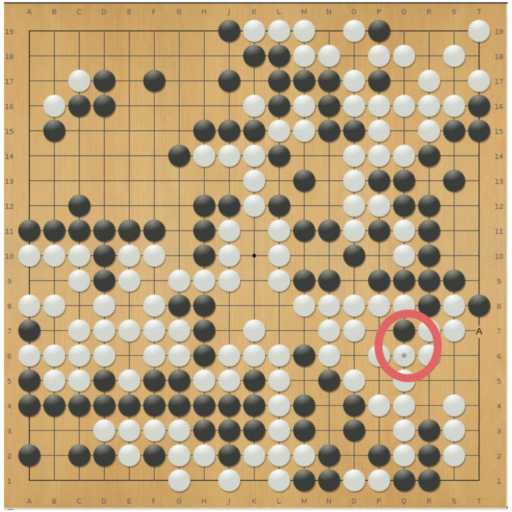
Other fun applications

(Xu et al., 2020)

win rate (black) >99.9%

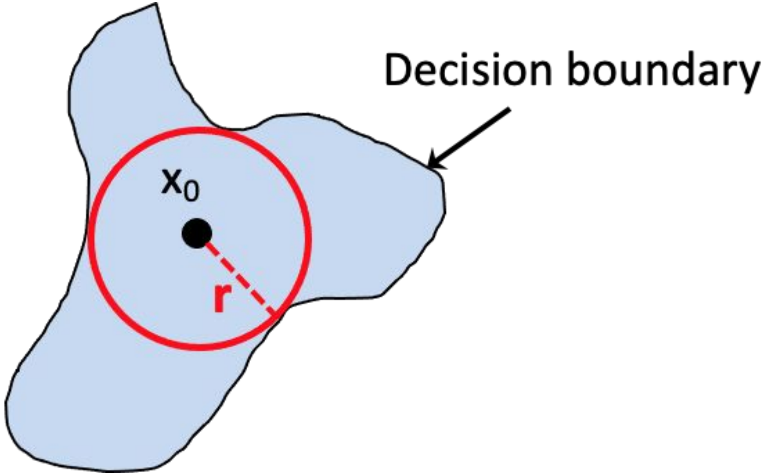


win rate (black) 10%



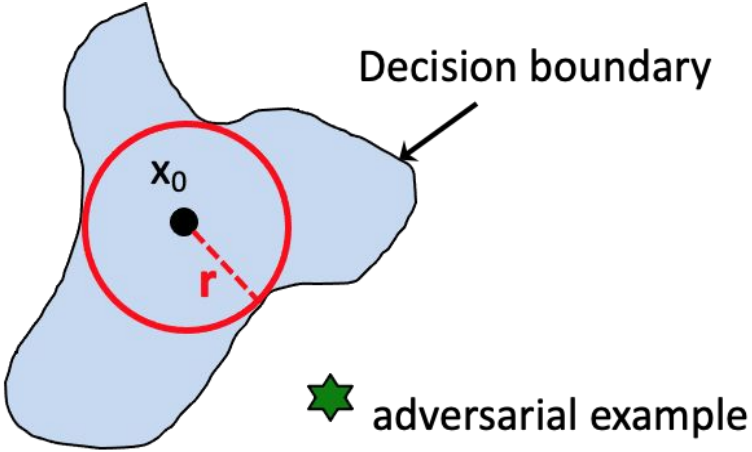
Mathematical definition

- Distance to the decision boundary (Lp norm or other metrics)
  - NP-hard for neural network with Lp norm (Katz et al., 2017)



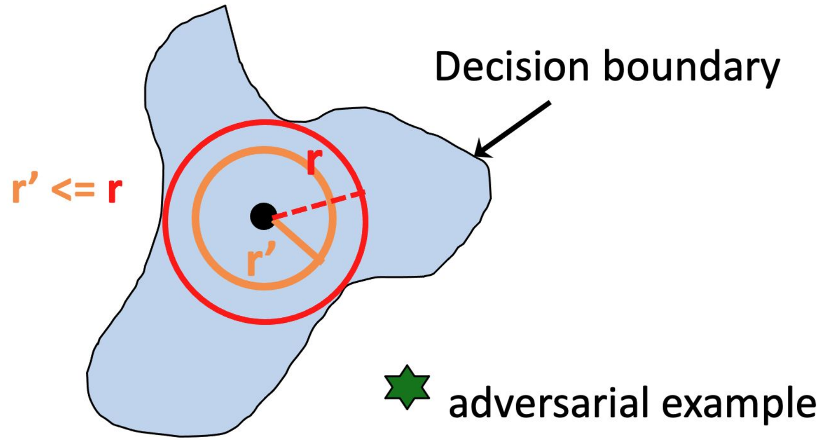
Mathematical definition

- Distance to the decision boundary
- Adversarial attack: over-estimation of robustness
  - White-box (Goodfellow et al., 2014, Carlini & Wagner, 2017, ...)
  - Black-box (Chen et al., 2017, Cheng et al., 2019,2020, ...)



Mathematical definition

- Distance to the decision boundary
- (Sound) verification: certified lower bound of robustness

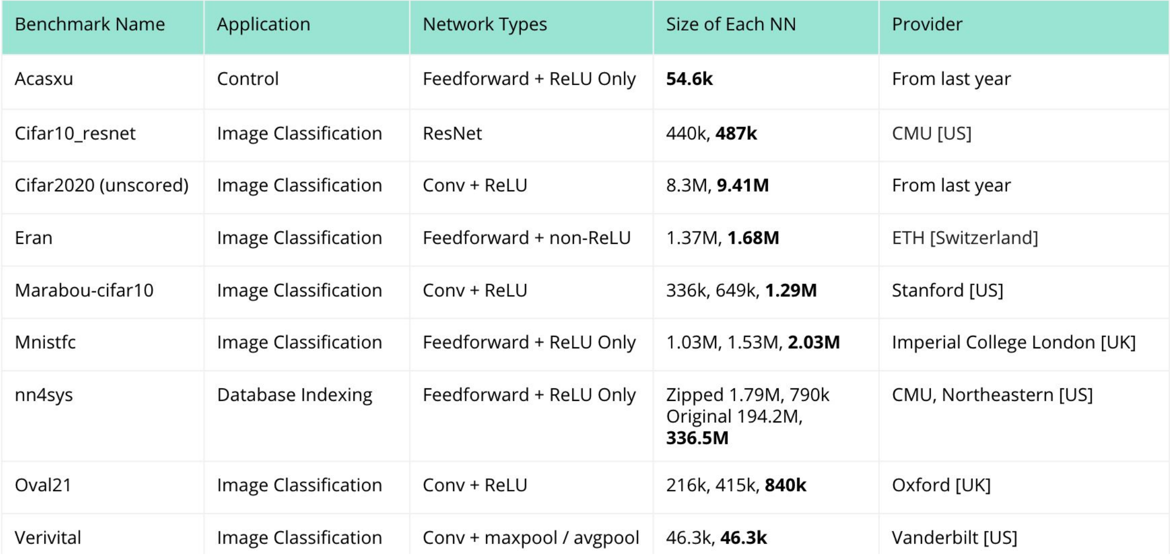


Challenges (attack)

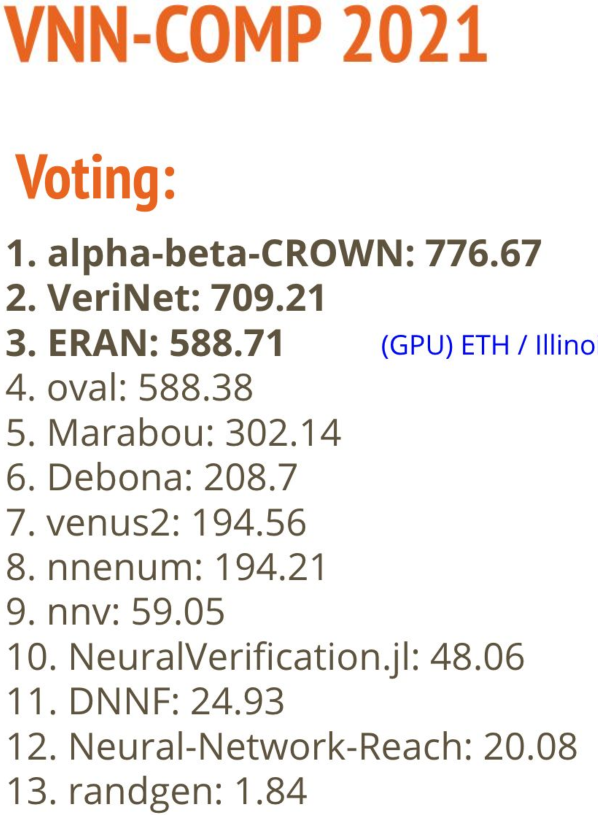
- Attack to real world systems?
  - Black box
  - Transferability
- Attacks beyond evasion

Challenges (verification)

- Reluplex (2017): ~100 neurons
- Alpha-Beta-Crown (2021): (semi-realistic) CIFAR models (>9M neurons)



Benchmark Name	Application	Network Types	Size of Each NN	Provider
Acasxu	Control	Feedforward + ReLU Only	**54.6k**	From last year
Cifar10_resnet	Image Classification	ResNet	440k, **487k**	CMU [US]
Cifar2020 (unscored)	Image Classification	Conv + ReLU	8.3M, **9.41M**	From last year
Eran	Image Classification	Feedforward + non-ReLU	1.37M, **1.68M**	ETH [Switzerland]
Marabou-cifar10	Image Classification	Conv + ReLU	336k, 649k, **1.29M**	Stanford [US]
Mnistfc	Image Classification	Feedforward + ReLU Only	1.03M, 1.53M, **2.03M**	Imperial College London [UK]
nn4sys	Database Indexing	Feedforward + ReLU Only	Zipped 1.79M, 790k Original 194.2M, **336.5M**	CMU, Northeastern [US]
Oval21	Image Classification	Conv + ReLU	216k, 415k, **840k**	Oxford [UK]
Verivital	Image Classification	Conv + maxpool / avgpool	46.3k, **46.3k**	Vanderbilt [US]



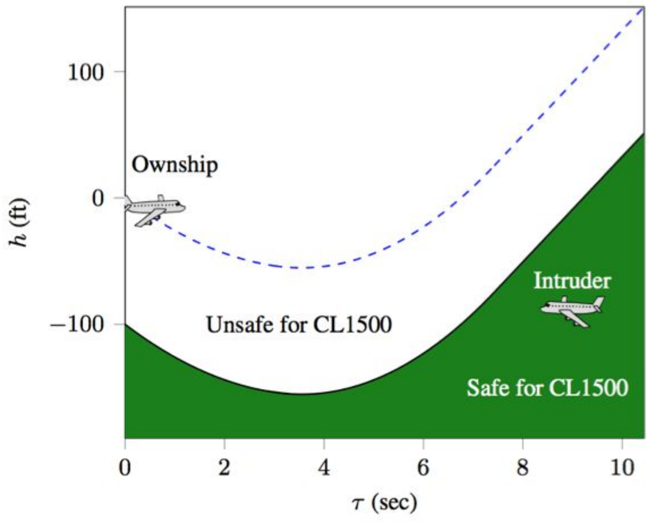
VNN-COMP 2021

Voting:

1. **alpha-beta-CROWN: 776.67**
2. **VeriNet: 709.21**
3. **ERAN: 588.71** (GPU) ETH / Illinois
4. oval: 588.38
5. Marabou: 302.14
6. Debona: 208.7
7. venus2: 194.56
8. nnenum: 194.21
9. nnv: 59.05
10. NeuralVerification.jl: 48.06
11. DNNF: 24.93
12. Neural-Network-Reach: 20.08
13. randgen: 1.84

Challenges (verification)

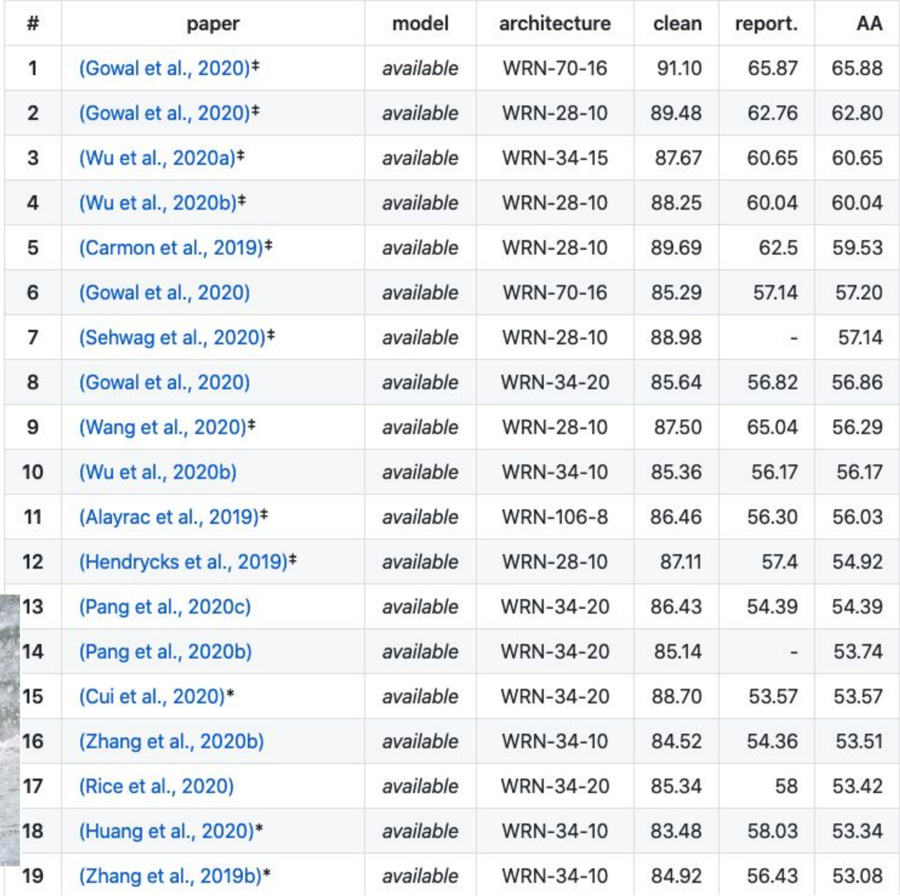
- Challenges:
  - **Lack of Real Applications**
  - “Hard” cases?
  - More realistic “specifications”?
  - More flexible architectures



Airborne collision avoidance system for drones (ACAS Xu)

Challenges (defense)

- Defense from real threads:
  - Maybe doesn't need to make NN robust?
  - Out-of-distribution data
  - Natural perturbations
- Robustness as regularization



#	paper	model	architecture	clean	report.	AA
1	(Gowal et al., 2020)‡	*available*	WRN-70-16	91.10	65.87	65.88
2	(Gowal et al., 2020)‡	*available*	WRN-28-10	89.48	62.76	62.80
3	(Wu et al., 2020a)‡	*available*	WRN-34-15	87.67	60.65	60.65
4	(Wu et al., 2020b)‡	*available*	WRN-28-10	88.25	60.04	60.04
5	(Carmon et al., 2019)‡	*available*	WRN-28-10	89.69	62.5	59.53
6	(Gowal et al., 2020)	*available*	WRN-70-16	85.29	57.14	57.20
7	(Sehwag et al., 2020)‡	*available*	WRN-28-10	88.98	-	57.14
8	(Gowal et al., 2020)	*available*	WRN-34-20	85.64	56.82	56.86
9	(Wang et al., 2020)‡	*available*	WRN-28-10	87.50	65.04	56.29
10	(Wu et al., 2020b)	*available*	WRN-34-10	85.36	56.17	56.17
11	(Alayrac et al., 2019)‡	*available*	WRN-106-8	86.46	56.30	56.03
12	(Hendrycks et al., 2019)‡	*available*	WRN-28-10	87.11	57.4	54.92
13	(Pang et al., 2020c)	*available*	WRN-34-20	86.43	54.39	54.39
14	(Pang et al., 2020b)	*available*	WRN-34-20	85.14	-	53.74
15	(Cui et al., 2020)*	*available*	WRN-34-20	88.70	53.57	53.57
16	(Zhang et al., 2020b)	*available*	WRN-34-10	84.52	54.36	53.51
17	(Rice et al., 2020)	*available*	WRN-34-20	85.34	58	53.42
18	(Huang et al., 2020)*	*available*	WRN-34-10	83.48	58.03	53.34
19	(Zhang et al., 2019b)*	*available*	WRN-34-10	84.92	56.43	53.08

Trustworthy Machine Learning: Challenges and Opportunities

Adversarial Machine Learning *for Good*

Pin-Yu Chen (IBM Research)
www.pinyuchen.com @pinyuchenTW

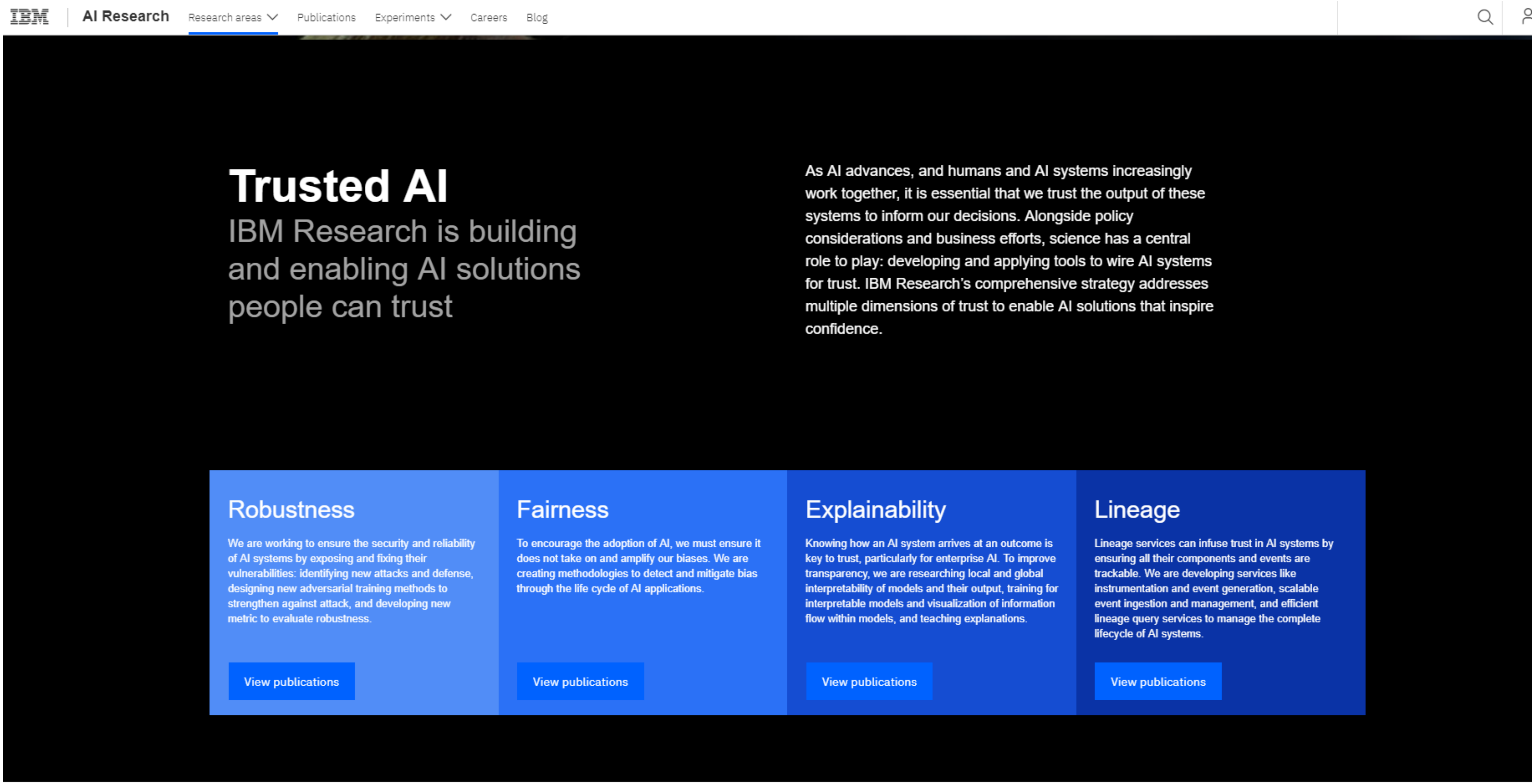
Machine Learning Summer School (MLSS@Taipei)
August 2021

IBM Research

The gap between AI development and deployment

How we develop AI

How we deploy AI



Trusted AI

IBM Research is building and enabling AI solutions people can trust

As AI advances, and humans and AI systems increasingly work together, it is essential that we trust the output of these systems to inform our decisions. Alongside policy considerations and business efforts, science has a central role to play: developing and applying tools to wire AI systems for trust. IBM Research's comprehensive strategy addresses multiple dimensions of trust to enable AI solutions that inspire confidence.

Robustness

We are working to ensure the security and reliability of AI systems by exposing and fixing their vulnerabilities: identifying new attacks and defense, designing new adversarial training methods to strengthen against attack, and developing new metric to evaluate robustness.

View publications

Fairness

To encourage the adoption of AI, we must ensure it does not take on and amplify our biases. We are creating methodologies to detect and mitigate bias through the life cycle of AI applications.

View publications

Explainability

Knowing how an AI system arrives at an outcome is key to trust, particularly for enterprise AI. To improve transparency, we are researching local and global interpretability of models and their output, training for interpretable models and visualization of information flow within models, and teaching explanations.

View publications

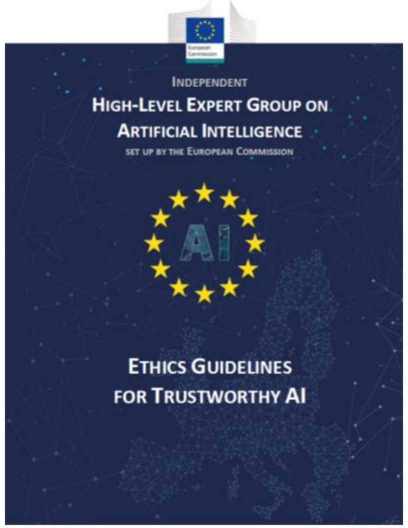
Lineage

Lineage services can infuse trust in AI systems by ensuring all their components and events are trackable. We are developing services like instrumentation and event generation, scalable event ingestion and management, and efficient lineage query services to manage the complete lifecycle of AI systems.

View publications

IBM Research AI

Definition of Trustworthy AI



INDEPENDENT
HIGH-LEVEL EXPERT GROUP ON
ARTIFICIAL INTELLIGENCE
SET UP BY THE EUROPEAN COMMISSION

ETHICS GUIDELINES
FOR TRUSTWORTHY AI

European Commission’s Definition

Trustworthy AI has **three components,** which should be met throughout the system's entire life cycle:

1. it should be **lawful**, complying with all applicable laws and regulations;
2. it should be **ethical**, ensuring adherence to ethical principles and values; and
3. it should be **robust**, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.

The General Data Protection Regulation (GDPR)

6. Integrity and confidentiality

Keep it secure

7. Accountability

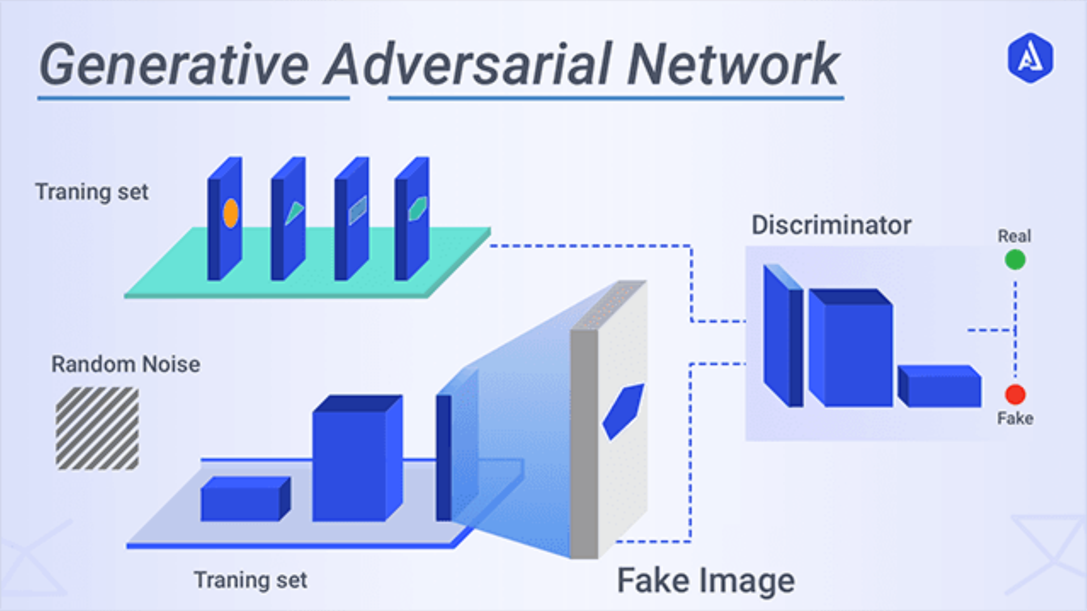
Record and prove compliance. Ensure policies.

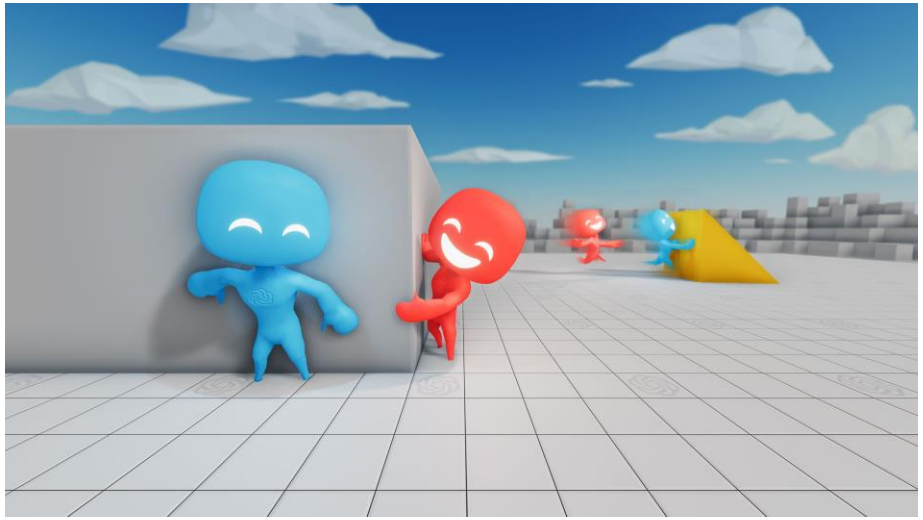
https://www.metacompliance.com/blog/what-are-the-7-principles-of-gdpr/

https://www.amara-marketing.com/travel-blog/7-principles-of-the-gdpr-and-what-they-mean

Adversarial ML: Learning with an Adversary

- Understanding model performance in the worst case
- Improving model performance by learning from (artificial) mistakes



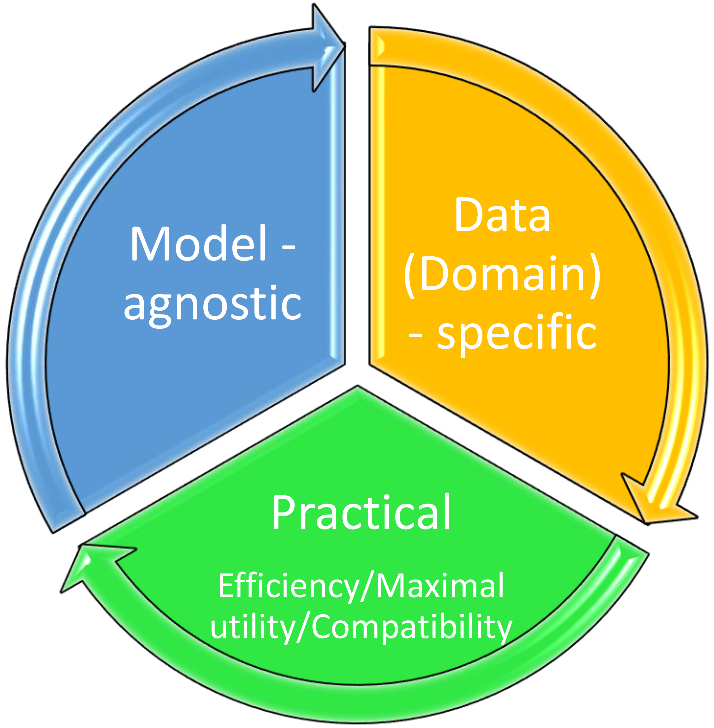


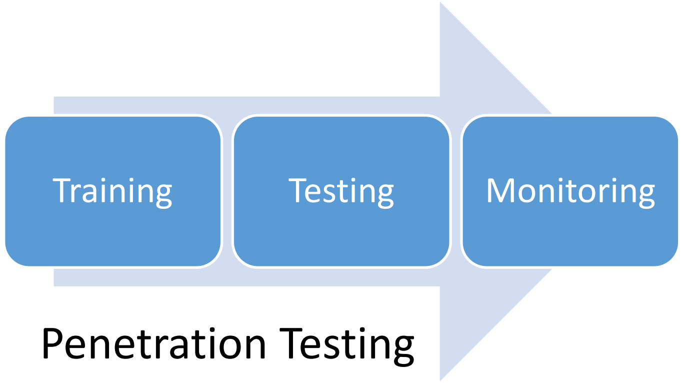
DEFENSE

Pentagon actively working to combat adversarial AI

https://www.akira.ai/glossary/generative-adversarial-networks/
https://openai.com/blog/emergent-tool-use/

Roadmap toward Holistic Adversarial Robustness





Attack (Bug Finding)

- In-house **sensitivity and reliability tests** for developed models
- Generate prediction-evasive examples (per user constraints)
- Customize to model deployment conditions (e.g. cloud APIs)

Defense (Model Hardening)

- **Detecting and mitigating** potential adversarial threats
- **Plug-and-play** model patching for a given model
- Landscape exploration: model fix and cleaning

Verification (Model Certificate)

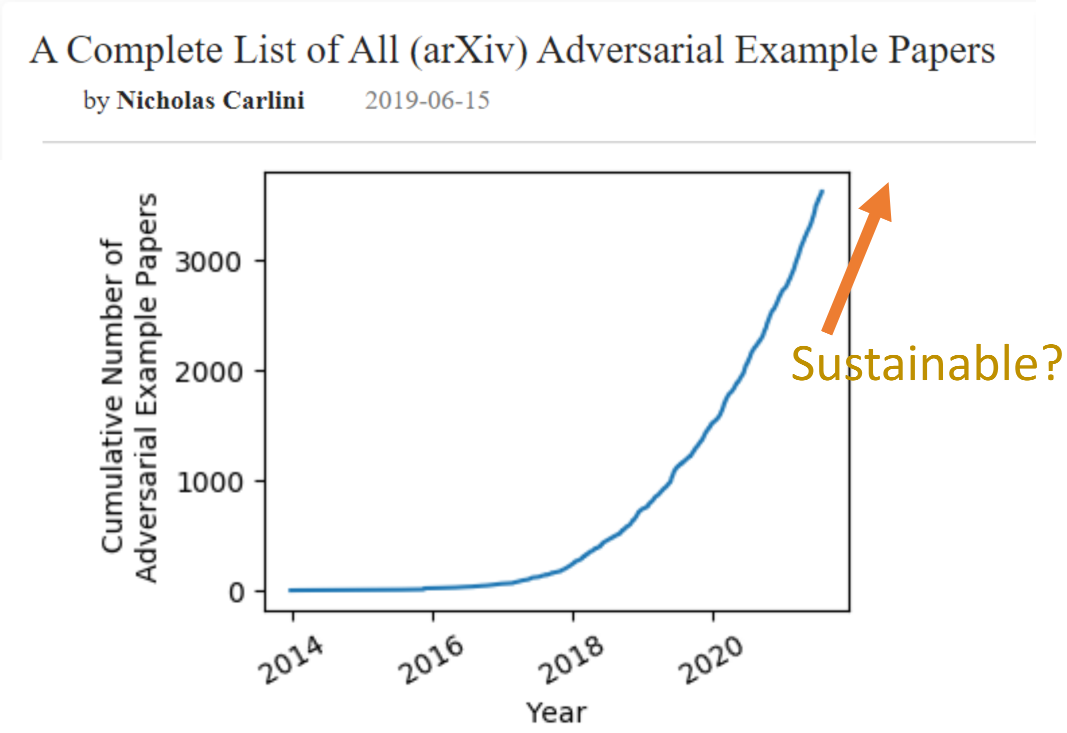
- This model is certified to be **attack-proof** up to a certain level
- Quantifiable metric for certified robustness
- AI standards, governance, and law regulation

Applications to AI (Model Boosting)

- Data augmentation
- **Model reprogramming**: data-efficient transfer learning
- Model watermarking

Trends I observed in Adversarial Machine Learning

- Attack:
  - Adversarial attack on [Task]
  - Black-box adversarial attack on [Task]
  - Hard-label black-box adversarial attack on [Task]
  - Efficient adversarial attack for [Perturbation Norm]
- Defense:
  - Defending against adversarial attacks using [Method]
  - Detecting adversarial examples using [Method]
  - Certified robustness for [Task]/[Norm]
  - Adversarial training using [Technique]
- Reflection:
  - All empirical defenses are vulnerable
  - How practical is the threat model? (e.g. unrestricted adversarial examples)
  - Intriguing properties of [New Network Architecture]
  - Tradeoff between adversarial robustness and [Factor] (e.g. privacy, fairness, interpretability)
  - Hardness of adversarial ML: optimization and generalization

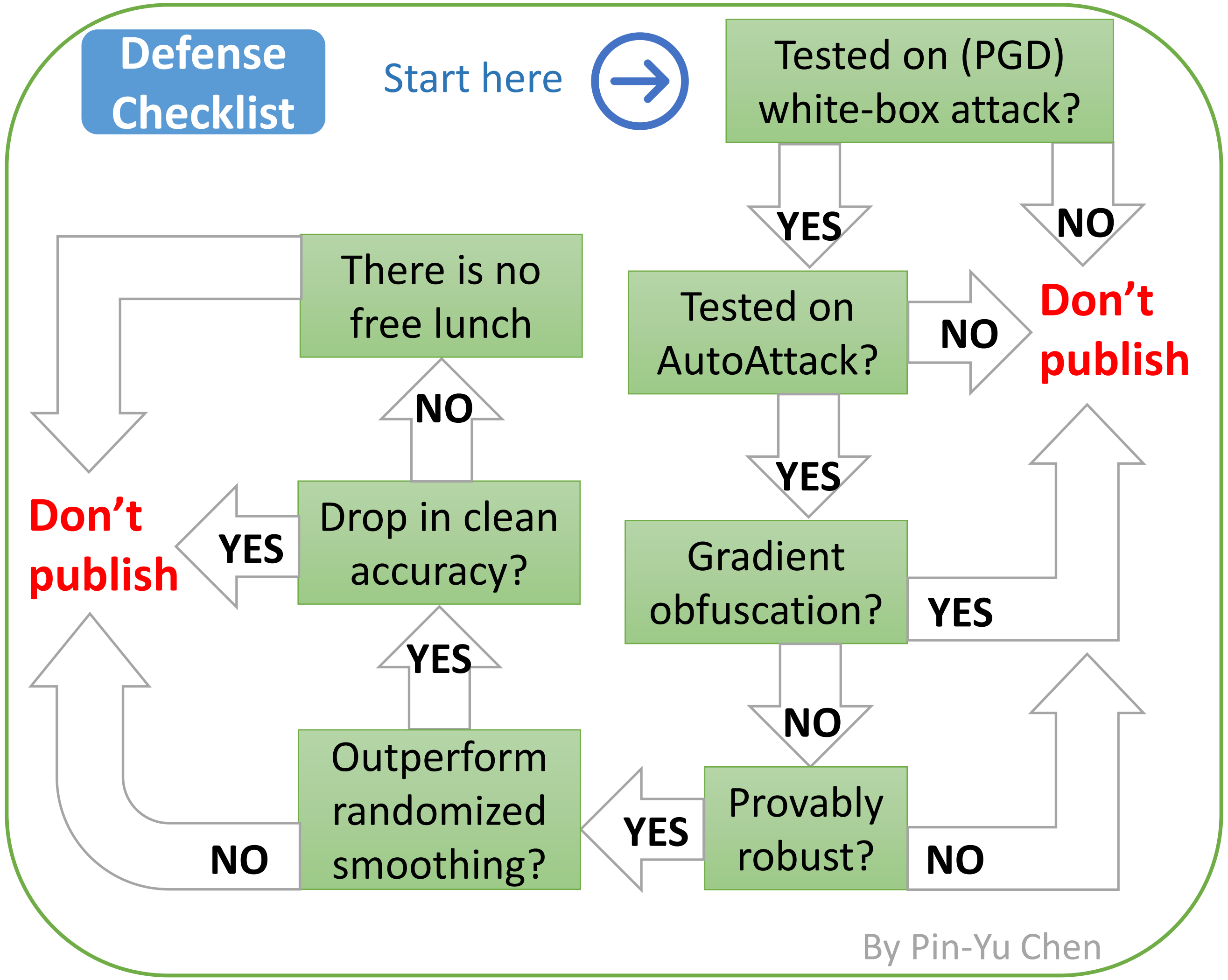


A Complete List of All (arXiv) Adversarial Example Papers

by **Nicholas Carlini** 2019-06-15

Sustainable?

Defense Checklist: Should I publish my defense against adversarial examples? [2021 version]



Panel Discussion –*Trustworthy Machine Learning: Challenges and Opportunities*

Sijia Liu, Assistant Professor,
Dept. Computer Science & Engineering,
Michigan State University

MLSS’21, TAIPEI

Short Bio

Trustworthy ML

- PhD, Syracuse University, 2011-2016
- Postdoc, University of Michigan, Ann Arbor, 2016-2017
- Research Staff Member, MIT-IBM Watson AI Lab, 2018-2020
- Assistant Professor, CSE, Michigan State University, 2021-

Optimization and Trustworthy AI Lab: https://lsjxjtu.github.io/

Today’s Three Focused Challenges:

1. (Attack) **Reverse Engineering of Deception (RED):** From Attack Generation, Rejection, to Attack Information Reverse Engineering

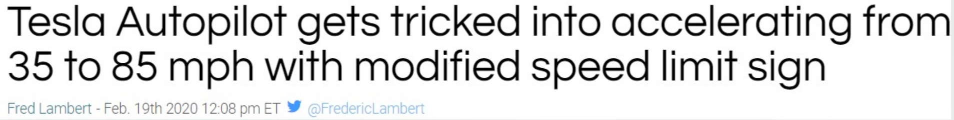
2. (Defense) **Algorithmic Foundation of Attack-Agnostic Defense**: Beyond Min-Max Adversarial Training

3. (System) **Robustness-to-X (R2X) Challenge**: Holistic view of robustness understanding

Attack Vision: From Generation, Rejection, to Reverse Engineering

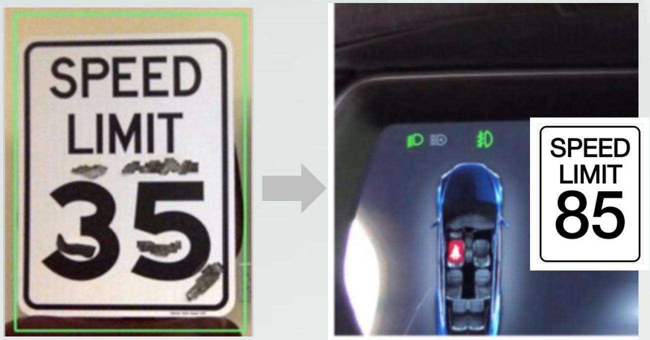
Adversarial attack: A standard way to evaluate ‘worst-case’ robustness of ML models

Existing work: Focuses on attack generation in diverse scenarios (digital/physical, white-box/black-box, soft/hard label, train-time/test-time)

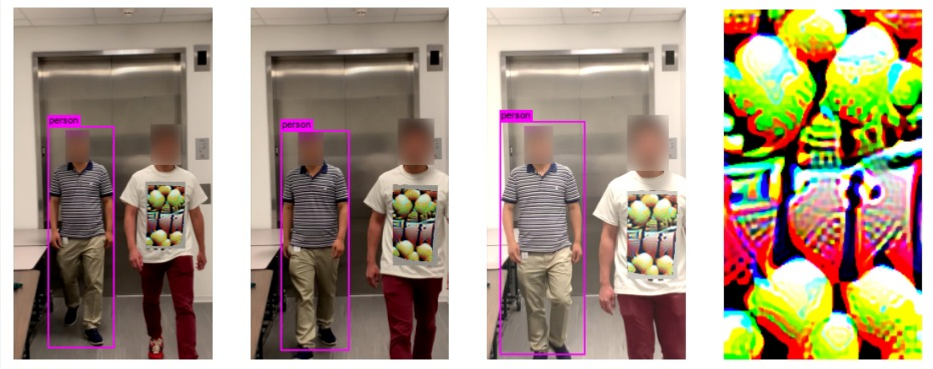


Tesla Autopilot gets tricked into accelerating from 35 to 85 mph with modified speed limit sign

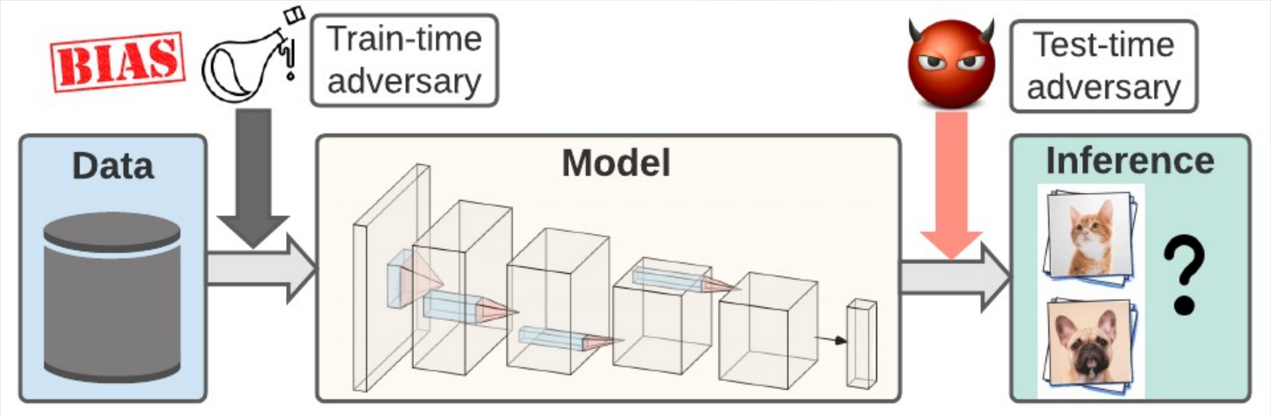
Fred Lambert - Feb. 19th 2020 12:08 pm ET @FredericLambert



Adversarial T-shirt to fool DNN-based person detectors [ECCV’20]



Attack Challenge: Reverse Engineering of Deception (RED)



RED aims to reverse engineer *attack toolchains*, rather than merely `rejecting' (in terms of detection or robust training) adversarial attacks.

1. **RED for train-time attack (backdoor/Trojan attack):** Recover *Trojan trigger* pattern given only Trojan model [ECCV'20, ICLR'21]

2. **RED for test-time attacks (adversarial examples)**: Recover *pixel-level perturbations* and *attribution-level attack saliency image region* from an attack [Feasibility and capability of RED?]

Defense Vision: From Attack-Specific Robust Training to Attack-Agnostic Robust Training

- **Min-max optimization based adversarial training** [ICLR'18]: Well-recognized algorithmic foundation for adversarial defense

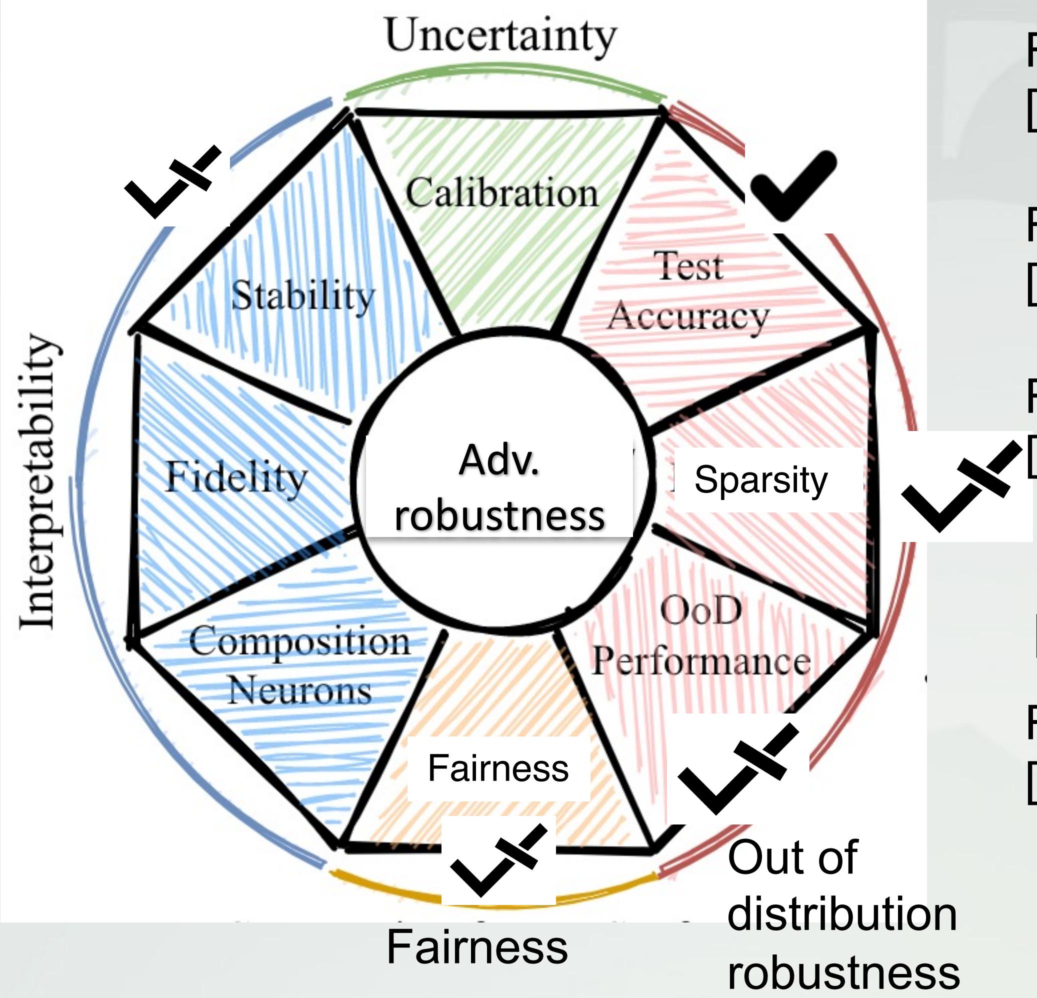
$$\underset{\boldsymbol{\theta}}{\text{minimize}}\ \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}}\left[\underset{\boldsymbol{\delta}\in\mathcal{C}}{\text{maximize}}\,\boxed{\ell_{\text{tr}}(\boldsymbol{\theta},\mathbf{x}+\boldsymbol{\delta},y)}\right]$$

- Attack-specific assumption: Attacker and defender share **same** objective, thus difficult to adapt to different types of attacks

- Attack-agnostic training: Attacker and defender would enjoy **different** objective functions --- Bi-Level Optimization (BLO)

$$\begin{array}{ll}\underset{\boldsymbol{\theta}}{\text{minimize}} & \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}}[\underline{\ell_{\text{tr}}(\boldsymbol{\theta},\mathbf{x}+\boldsymbol{\delta}^*(\boldsymbol{\theta};\mathbf{x},y),y)}] \\ \text{subject to} & \boldsymbol{\delta}^*(\boldsymbol{\theta};\mathbf{x},y)=\underset{\boldsymbol{\delta}\in\mathcal{C}}{\arg\min}\,\underline{\ell_{\text{atk}}(\boldsymbol{\theta},\boldsymbol{\delta};\mathbf{x},y)}\end{array}$$

A Holistic View: Robustness-to-X (R2X)



Robustness vs. accuracy: e.g., [ECCV’18, ICML’19, ICLR’19]

Robustness vs. sparsity: e.g., [ICCV’19]

Robustness vs. OOD: e.g., [ECCV’20, NeurIPS’20]

Robustness vs. fairness: e.g., [ICML’21, FAT’21]

Robustness vs. interpretability: e.g., [NeurIPS’19, ICML’20]

"Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning."

~ Winston Churchill