SEI Thoughts on AI T&E and Related Topics

APRIL 13, 2023

Dr. Tom Longstaff
CTO
CMU Software Engineering Institute

Carnegie Mellon University
Software Engineering Institute

Legal

SEI Insights on T&E

Carnegie Mellon University
Software Engineering Institute

SEI Insights on T&E

Carnegie Mellon University Software Engineering Institute

Dev/D&OT&E/Sustainment Integration

- We need to consider an integrated infrastructure to connect all portions of the lifecycle to achieve continuous V&V/ATO.
- Changing conditions in deployed environments are driving to incremental and continuous software (and hardware) modifications after the initial ATO.

Culture Change

Just saying we need to change the culture is not enough. There are some underlying premises and constructs that need to be rigorously addressed. The Operational Test culture reflects the mission, which is to conduct rigorous, ***objective*** T&E under ***operational*** conditions. How is this mission maintained in a shift-left, shared data environment?

- How do you shift left and still be objective?
- How can data from shifted-left environments be used to justify operational release?

SEI Insights on T&E (continued)

Carnegie Mellon University Software Engineering Institute

Integrate Policy & Guidance Development

Integrate DT and OT equities directly into P&G development that affects them (e.g., DoD CIO DSO guidance, SW Acquisitions Pathway guidance). For example, code maturity is a "thing" for DT and OT but not in DSO CIO DSO guidance. DOT&E should be more closely involved in a review and update to the DoD CIO DSO guidance on the SW Pathway.

DoD Pipeline Ownership

What role if any does DT and OT play in pipeline tool selection, pipeline development and pipeline testing?

Does the role of DT and OT differ if the government owns the pipeline as opposed to the contractor?

DoD pipeline ownership implies that the government is implicated in the results. This effects underlying premises about what is and what is not DT's and OT's "business."

Specific DT and OT concern with Agile

How are the product vision and the needs of the overall stakeholder community protected in an agile development environment centered on specific customers? The drift from legacy systems has downstream implications. Recommend a more strategic approach to the integration of agile updates to mission systems.

SEI Insights on T&E

Carnegie Mellon University Software Engineering Institute

Digital Engineering

Siloed and locked-in models are preventing the adoption of digital engineering. Models need to be developed iteratively with iterative and incremental execution of threads across models.

We don't want to go back to waterfall "big design upfront" only with models instead of code. Assure models have the same shift-left and agility as code.

The old assumption was that physical reality is the gold standard for testing. Is that the case when you have instances of physical test being executed to provide data to the model? Models in some cases provide a more accurate reflection of potential reality than what can be achieved on a physical range.

Physical computational limits impact the ability to effectively model operational reality.

We need mechanisms to V&V models to give DT and OT confidence. Transparency and evidence regarding the models' reflection of the operational environments is required.

Models of user behavior will help obviate the extent of real-world user testing.

Humans in the loop: Models must be consumable and interact-able by humans, not just by other models and tools. Maybe at some point the models will interact with AI bots and humans won't be that much in the picture, but first put humans in the loop and then iteratively figure out where they can be replaced.

Plug & play models: We need to have standards to enable interaction by models from multiple vendors by building them to common environmental interfaces.

TESTING, EVALUATING, AND ASSESSING ARTIFICIAL INTELLIGENCE-ENABLED SYSTEMS UNDER OPERATIONAL CONDITIONS FOR THE DEPARTMENT OF THE AIR FORCE

Air Force Studies Board

NATIONAL ACADEMIES Sciences Engineering Medicine

Study Origin and Output

Study requested by the USAF 96th Test Wing

Air Force Materiel Command referred the study to AFSB

Proceedings of a workshop—in brief

Consensus study report

Study Tasks

“Examine the Air Force Test Center's technical capabilities and capacity to conduct rigorous and objective test, evaluation, and assessments of artificial intelligence (AI)-enabled systems under operational conditions and against realistic threats. Specifically, the committee will

1) Evaluate and contrast current testing and assessment methods employed by the Department of the Air Force and in commercial industry.
2) Consider examples of AI corruption under operational conditions and against malicious cyber attacks.
3) Recommend promising areas of science and technology that may lead to improved detection and mitigation of AI corruption.

Study Committee

- May Casterline, NVIDIA (co-chair)
- Tom Longstaff, Software Engineering Institute, Carnegie Mellon (co-chair)
- Brig Gen Craig Baker, USAF (ret)
- Bob Bond, MIT Lincoln Laboratory
- Rama Chellappa, Johns Hopkins University
- Trevor Darrell, University of California-Berkeley
- Melvin Greer, Intel Corporation
- Tammy Kolda (NAE), mathsci.ai
- Nandi Leslie, Raytheon Technologies
- Robin Murphy, Texas A&M University
- David Rosenblum, George Mason University
- Lt Gen (ret) Jack Shanahan, USAF
- Humberto “Tito” Silva, Sandia National Laboratories
- Rebecca Willett, University of Chicago

Project Schedule

April 22, 2022:	Kickoff meeting
June 27-29, 2022:	Data-gathering workshop
August 22, 2022:	Data-gathering meeting
September 28-29:	Data-gathering and writing meeting
Fall/winter:	Data-gathering and writing meetings
Early 2023:	Finalize report and go into review
Mid 2023:	Report release

Data-Gathering Workshop

- Eileen Bjorkman, Air Force Test Center
- Jane Pinelis, CDAO (Office of the DoD Chief Digital and Artificial Intelligence Officer)
- Chad Bieber, CDAO
- Marshall Kendrick, 45th Test Squadron, USAF
- David Coppler, 46th Test Squadron, USAF
- Jacob Martinez, 47th Cyberspace Test Squadron, USAF
- Olivia Brown, MIT LL
- Michael Wellman, University of Michigan
- Tom Strat, DZYNE Technologies
- Jim Bellingham, Johns Hopkins University
- Nancy Cooke, Arizona State University
- Bin Yu, UC Berkeley
- Nathan VanHoudnos, SEI
- Matt Turek, DARPA
- Bruce Draper, DARPA
- Ed Zelnio, AFRL

What's Next?

- More data-gathering sessions to come
- Workshop proceedings in brief coming soon
- Transforming the rough outline into a published report

The Influence of AI Engineering on T&E

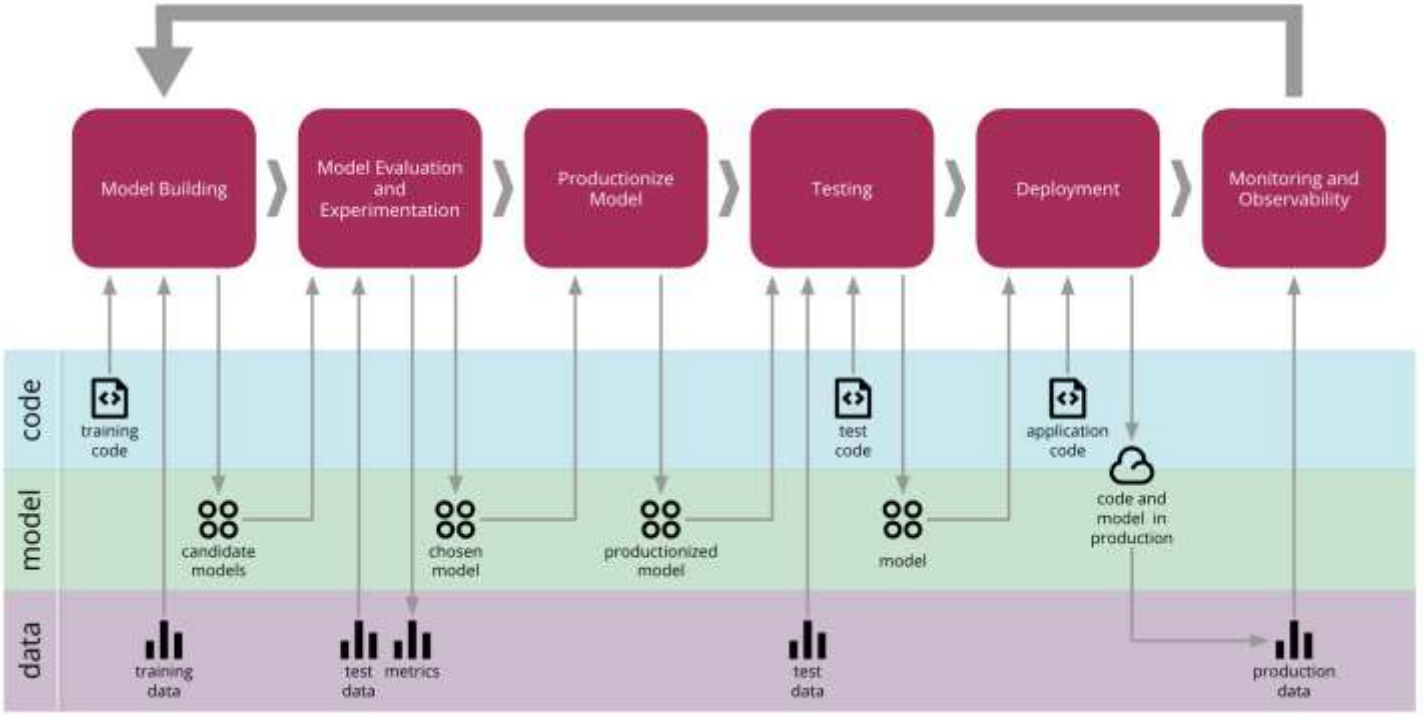
Carnegie Mellon University
Software Engineering Institute

AI Engineering Pillars

Carnegie Mellon University Software Engineering Institute

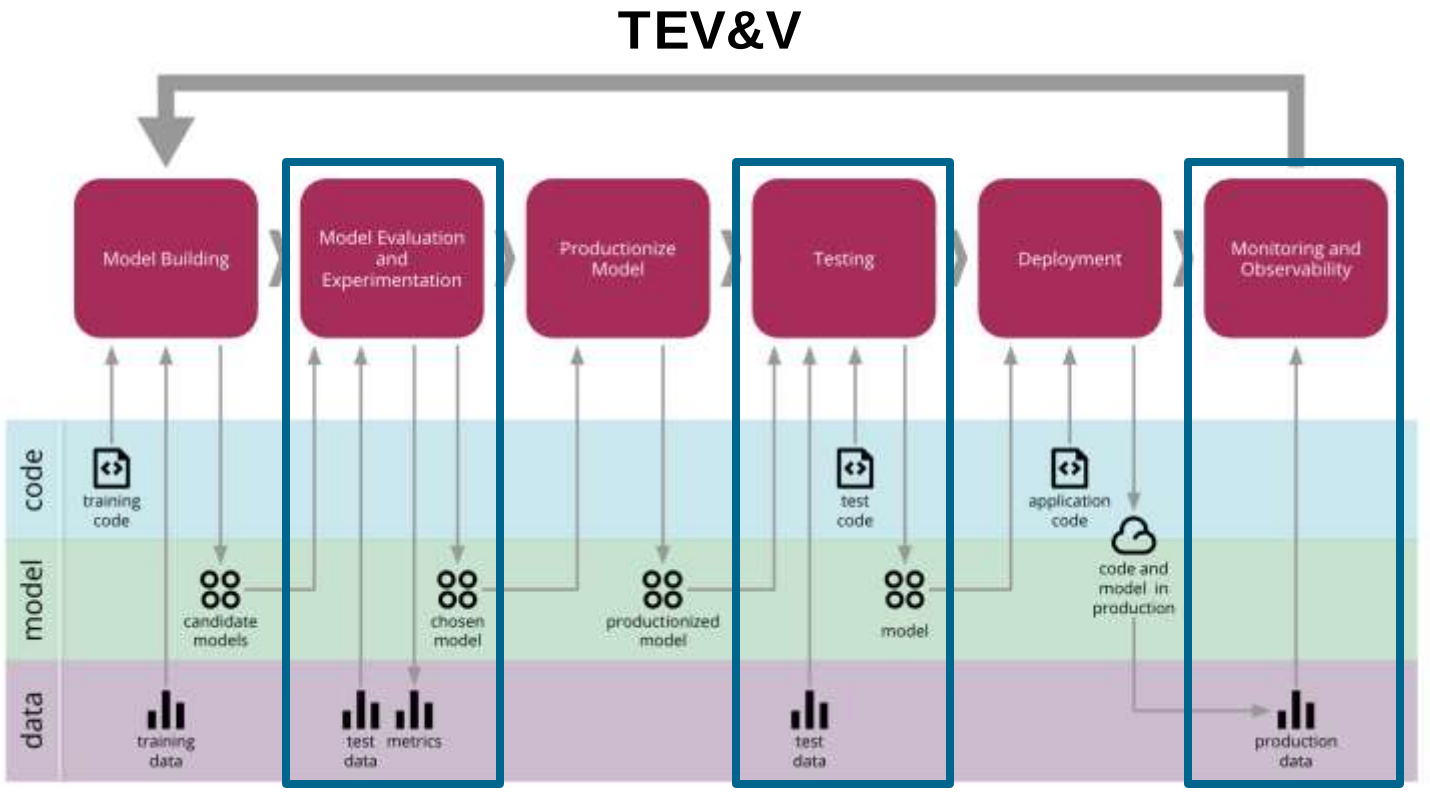
	**Human-Centered AI** *Design with the goal of working with, and for, people*	• Understand context of use and sense changes over time • Scope and facilitate human-machine teaming • Methods, mechanisms, and mindsets for critical oversight
	**Robust and Secure AI** *Operate reliably when faced with uncertainty or threat*	• Robustness of AI components and systems • Designing for security challenges in modern AI systems • Testing, evaluating, and analyzing AI systems
	**Scalable AI** *Accommodate the size, speed, and complexity of mission needs*	• Scalable management of data and models • Enterprise scalability of AI development and deployment • Scalable algorithms and infrastructure

Getting ML Models into Production



Source: Continuous Delivery for Machine Learning, Martin Fowler

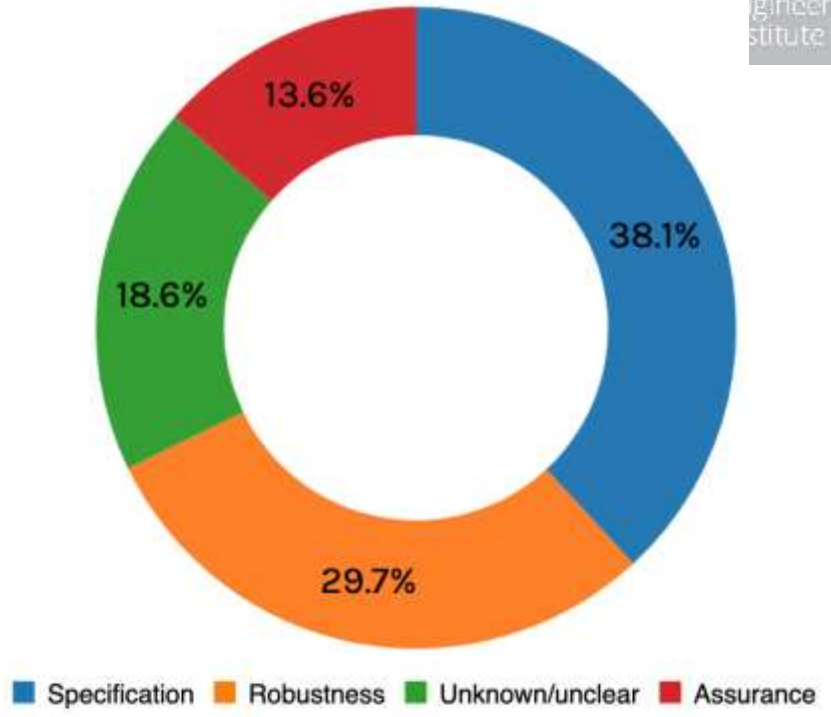
Getting ML Models into Production



Source: [Continuous Delivery for Machine Learning, Martin Fowler]

AI Engineering: Why Do AI Systems Fail?

Specification: The system's behavior did not align with the true intentions of its designer, operator, etc.
Robustness: The system operated unsafely because of features or changes in its environment, or in the inputs the system received.
Assurance: The system could not be adequately monitored or controlled during operation.



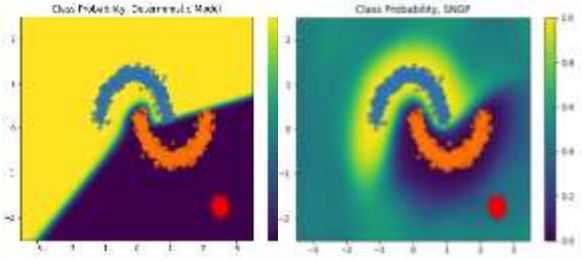
Source: https://incidentdatabase.ai/taxonomy/cset Retrieved 3/7/2023.
Credit to Partnership on AI and the Center for Security and Emerging Technologies (CSET) at Georgetown University.

"Failure is central to engineering. Every single calculation that an engineer makes is a failure calculation. Successful engineering is all about understanding how things break or fail."
– Henry Petroski

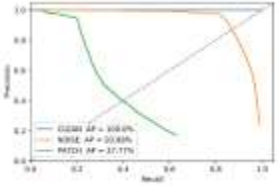
Beyond Accuracy

CSWaP constraints

Uncertainty quantification



Security/Adversarial Robustness

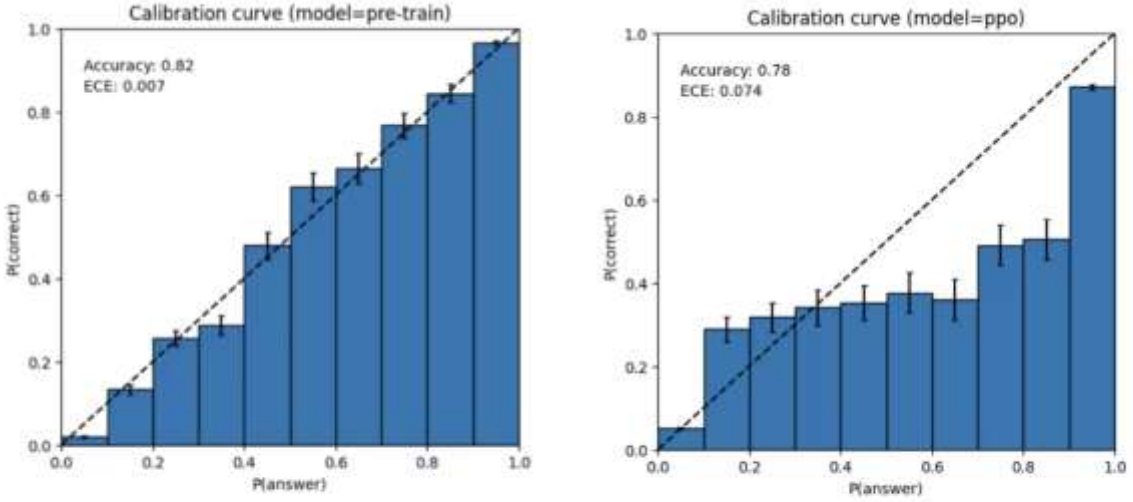


(Adhikari et al., 2020)

Need to understand the **trade-offs**

- task accuracy
- business/mission case
- robustness
- computational cost of training
- computational cost of inference
- deployment form factor (CSWaP)
- risk/threat/resilience/harms
- interpretability/explainability
- responsible AI
- …

Recent Observation: GPT-4 and Calibrated



Left: Calibration plot of the pre-trained GPT-4 model on an MMLU subset. The model's confidence in its prediction closely matches the probability of being correct. The dotted diagonal line represents perfect calibration. Right: Calibration plot of post-trained PPO GPT-4 model on the same MMLU subset. Our current process hurts the calibration quite a bit.

"GPT-4 can also be confidently wrong in its predictions, not taking care to double-check work when it's likely to make a mistake. Interestingly, the base pre-trained model is highly calibrated (its predicted confidence in an answer generally matches the probability of being correct). However, through our current post-training process, the calibration is reduced."

March 14, 2023: https://openai.com/research/gpt-4

The Advanced Computing Lab—AI Division

ADVANCED COMPUTING LAB

Test and Evaluation

DARPA Domain-Specific System on Chip (DSSoC)
DARPA Data Protection in Virtual Environments (DPRIVE)
DARPA Software Defined Hardware (SDH)

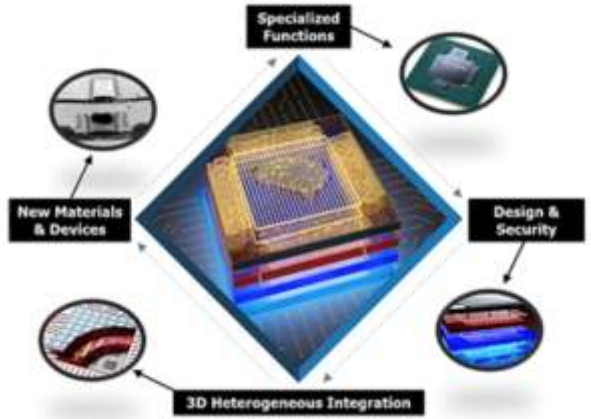
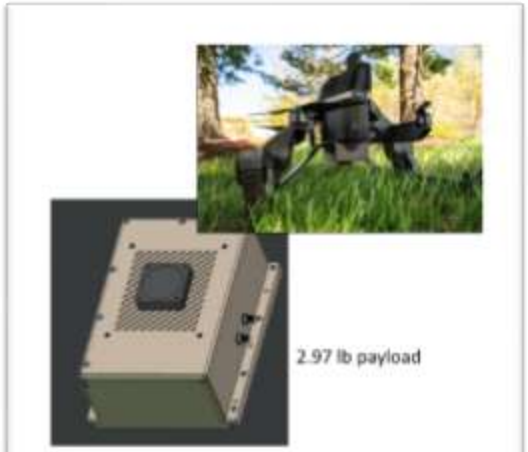


Image: DARPA

Internal Research

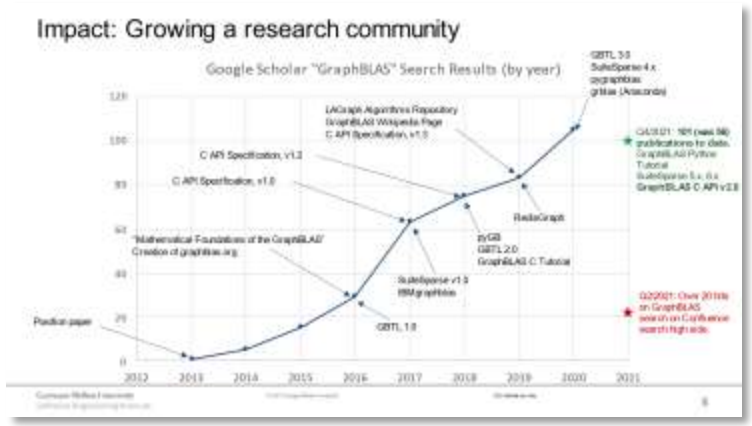
Portable, High-performance Inference at the Tactical Edge (PHITE)
Co-Design for Edge AI
Spiral AI/ML
Spiral Graph



UAS with Prototy pe STORM Payload

Technical Advisory

Test, evaluate, monitor, and provide expertise for entity developing a GraphBLAS implementation
Advising GraphBLAS implementations on new hardware



SEI Presentation to the NAS Study on AI T&E

Carnegie Mellon University
Software Engineering Institute

Study Goals Addressed in This Talk

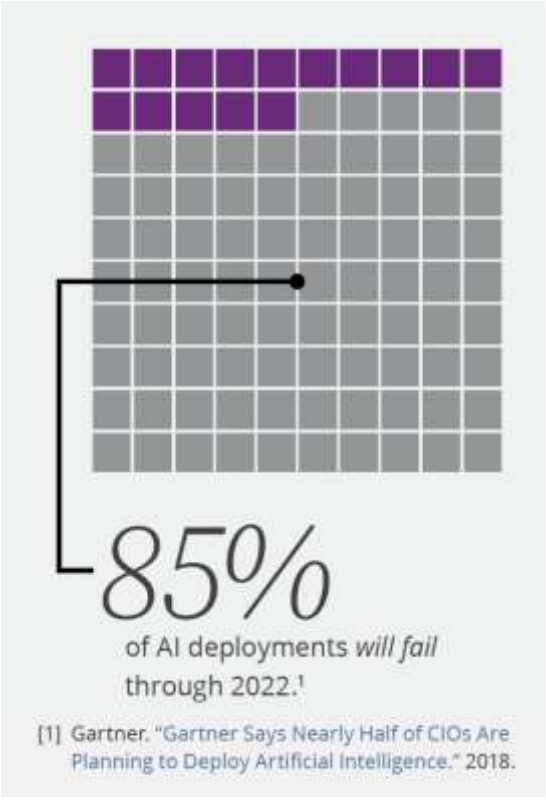
Testing, Evaluating, and Assessing Artificial Intelligence-Enabled Systems under Operational Conditions for the Department of the Air Force

Specifically, the committee will

1. Evaluate and contrast current testing and assessment methods employed by the Department of the Air Force and in commercial industry.
2. Consider examples of AI corruption under operational conditions and against malicious cyber attacks.
3. Recommend promising areas of science and technology that may lead to improved detection and mitigation of AI corruption.

Agenda: Introduction / Goal 2 / Goal 3

AI Engineering at the Software Engineering Institute



AI Engineering

- field of research and practice
- integrates software engineering, systems, CS, and human-centered design
- builds AI responsive to human needs and mission outcomes.

Human-Centered

Works with and for people

Scalable

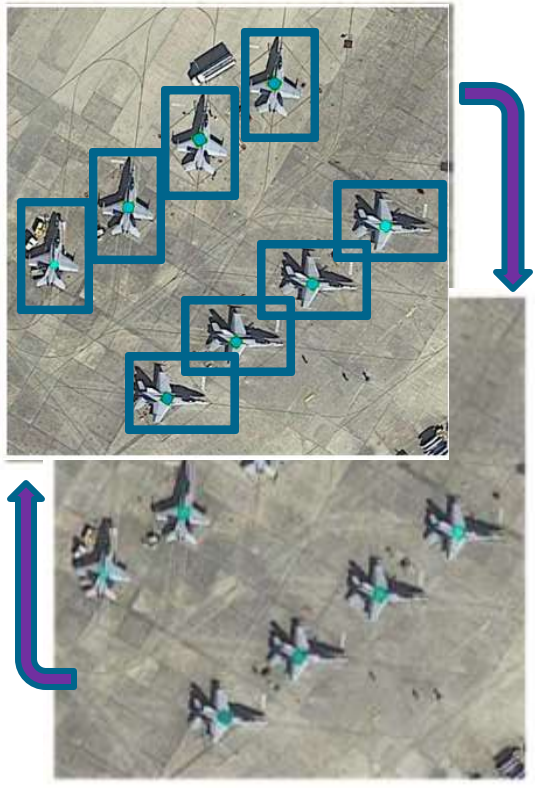
Size, speed, & complexity of mission needs

Robust and Secure

Reliable when under uncertainty or threat

https://sei.cmu.edu/our-work/artificial-intelligence-engineering

AML Lab: Applied Research Lab for Security of AI

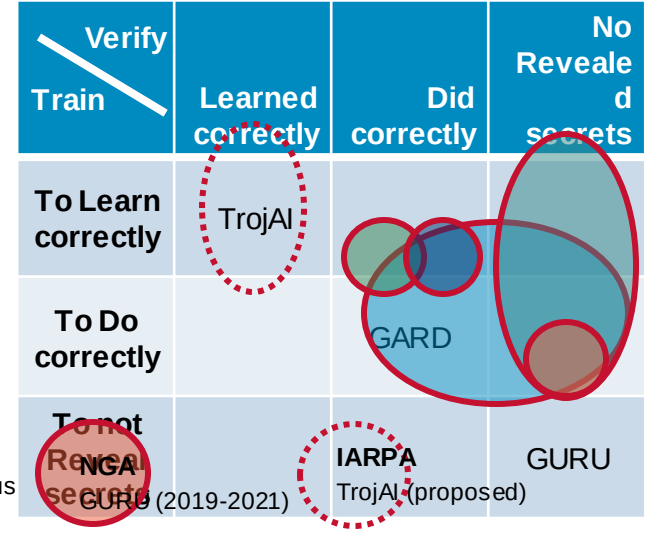


Source of images:
Adhikari et al. (2020)

SEI	Make software as good as it can be for the DoD.
AI Division	Make AI as good as it can be for the DoD.
AML Lab	Make AI as secure as it can be for the DoD.

"Secure" = secure across Beieler (2019) taxonomy:
An adversary can make you

- Learn the wrong thing
- Do the wrong thing
- Reveal the wrong thing



Study Goals Addressed in This Talk

Testing, Evaluating, and Assessing Artificial Intelligence-Enabled Systems under Operational Conditions for the Department of the Air Force

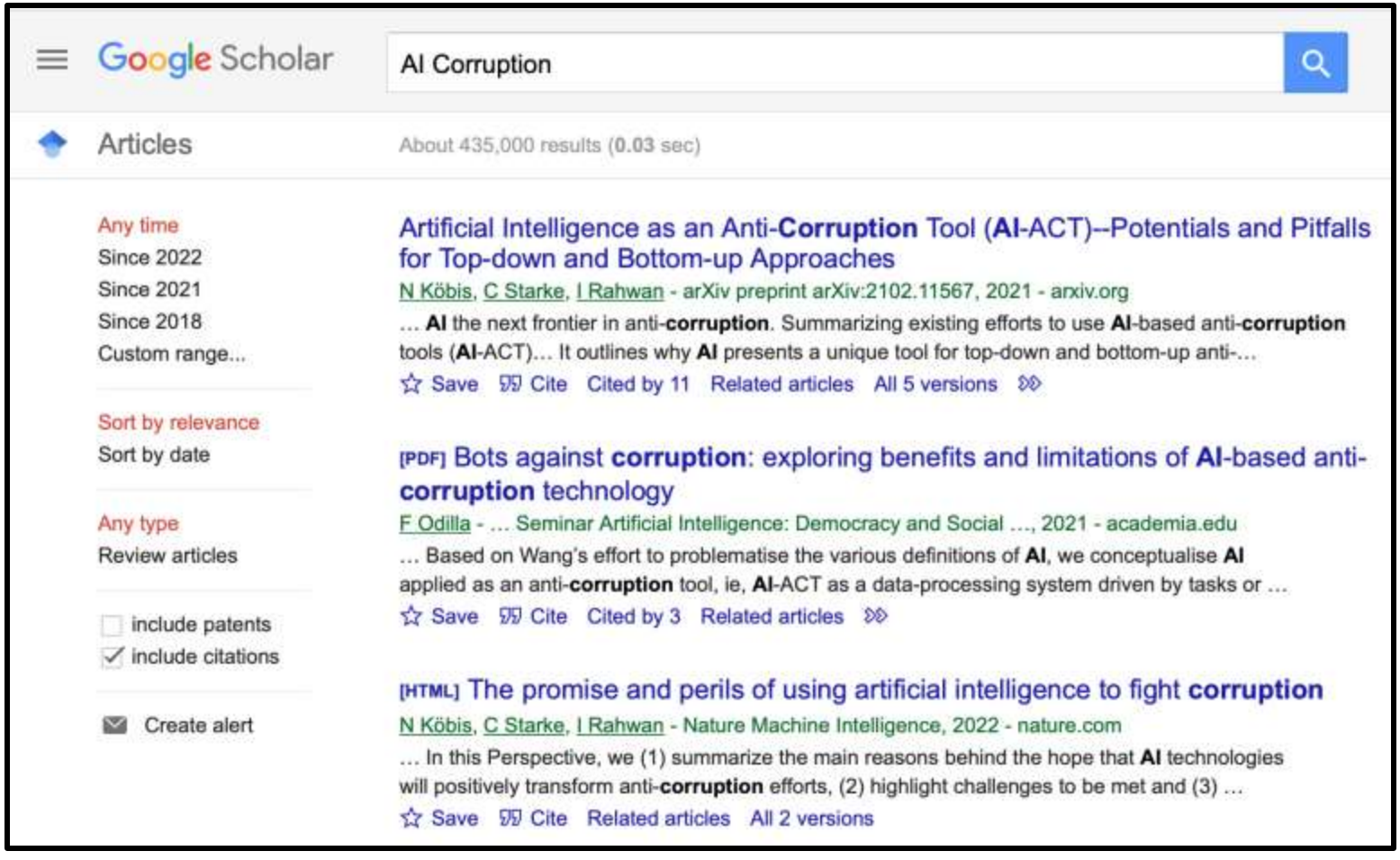
Specifically, the committee will:

1. Evaluate and contrast current testing and assessment methods employed by the Department of the Air Force and in commercial industry.
2. Consider examples of AI corruption under operational conditions and against malicious cyber attacks.
3. Recommend promising areas of science and technology that may lead to improved detection and mitigation of AI corruption.

Agenda: Introduction / Goal 2 / Goal 3

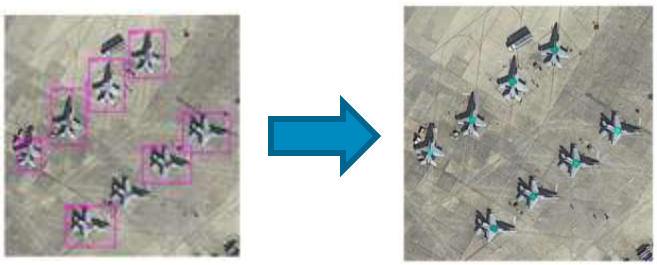
Goal 2: Defining AI Corruption / An Adversarial Perspective on Testing

AI Corruption Is Not Yet a Term of Art



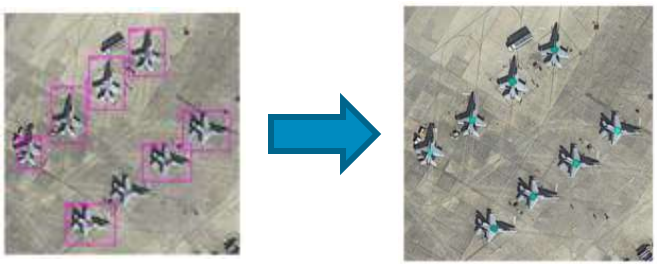
AI Corruption: A Decrease in a Quality Attribute of an AI System

Domain	Computer Vision	Computer Vision
ML Task	Object Detection	Object Detection
Mission	Broad Area Surveillance	Reconnaissance
Quality Attribute	High recall of target objects	High precision
AI Corruption	Recall lowered by physical patterns (Adhikari et al., 2020)	Precision lowered by environmental conditions (glint from snow)



AI Corruption Applies to Counter-AI systems

Domain	Computer Vision	Computer Vision
ML Task	Object Detection	Counter Object Detection
Mission	Broad Area Surveillance	Denial and Deception
Quality Attribute	High recall of target objects	Low recall of target detector
AI Corruption	Recall lowered by physical patterns (Adhikari et al., 2020)	AI Assurance defenses deployed by target detector to raise recall (Chou et al, 2020)



Study Goals Addressed in This Talk

Testing, Evaluating, and Assessing Artificial Intelligence-Enabled Systems under Operational Conditions for the Department of the Air Force

Specifically, the committee will:

1. Evaluate and contrast current testing and assessment methods employed by the Department of the Air Force and in commercial industry.
2. Consider examples of AI corruption under operational conditions and against malicious cyber attacks.
3. Recommend promising areas of science and technology that may lead to improved detection and mitigation of AI corruption.

Agenda: Introduction / Goal 2 / Goal 3

Goal 2: Defining AI Corruption / An Adversarial Perspective on Testing

Adversarial Testing Focuses on Risk, Metrics, and Complexity

Measuring AI Systems Beyond Accuracy

Violet Turri[1], Rachel Dzombak[2], Eric Heim[3], Nathan VanHoudnos[4], Jay Palat[5], Anusha Sinha[6]
Carnegie Mellon University Software Engineering Institute
vturri@sei.cmu.edu[1], rdzombak@sei.cmu.edu[2], eheim@sei.cmu.edu[3], nvanhoudnos@sei.cmu.edu[4],
rpalat@sei.cmu.edu[5], asinha@sei.cmu.edu[6]

Abstract

Current test and evaluation (T&E) methods for assessing machine learning (ML) system performance often rely on incomplete metrics. Testing is additionally often siloed from the other phases of the ML system lifecycle. Research investigating cross-domain approaches to ML T&E is needed to drive the state of the art forward and to build an Artificial Intelligence (AI) engineering discipline. This paper advocates for a robust, integrated approach to testing by outlining six key questions for guiding a holistic T&E strategy.

Introduction

Most machine learning projects focus on "accuracy" for model evaluation. While accuracy is useful for knowing how well a model performs on a test dataset at the time of model development, there are other significant implications in assessing the utility and usability of a machine learning model. Key considerations include robustness, resilience, calibration, confidence, alignment with evolving user requirements, and fit for mission and stakeholder needs as part of an integrated system, among others. In this paper, we explore what it means to measure beyond accuracy and define critical considerations for the test and evaluation of machine learning and, more broadly, artificial intelligence systems. After defining six key considerations related to robust T&E, the AI engineering community will be better equipped to develop and implement comprehensive and applied methods for the evaluation of models as well as possible metrics for more realistic and real-world model evaluation.

Current AI T&E Practices

Modern AI systems, many of which are built using machine learning, are a departure from static software systems that yield deterministic results. In contrast to analytical systems that follow explicit "instructions" given by a programmer and can be reduced and decomposed, AI systems are empirical, opaque, and unpredictable — they behave based on what they "learn" from data or experience (Russell & Norvig, 2021). Because AI systems often model real-world relationships, they must be adaptable to changing inputs and shifting correlations.

The characteristic differences between traditional software and AI lead to a series of open questions around the design and implementation of AI. For example, AI systems can evolve and change behavior over time. How can we ensure that AI systems are still doing what they are supposed to do? How can we certify that they are safe and reliable? Many current AI and ML methods are data intensive; continuous updates to data, while necessary, can impact architectural concerns such as prediction accuracy and latency (Ozkaya, 2020). Technical debt, driven by data dependencies, accumulates rapidly, silently, and at the system level (Sculley et al., 2015). What practical mechanisms exist to evaluate the state of a system when using large and evolving data sets? Further, exhaustive testing is currently not possible for systems that learn and adapt. How do we change approaches to test and evaluation to be more risk-, resilience-, and process- focused rather than exhaustive?

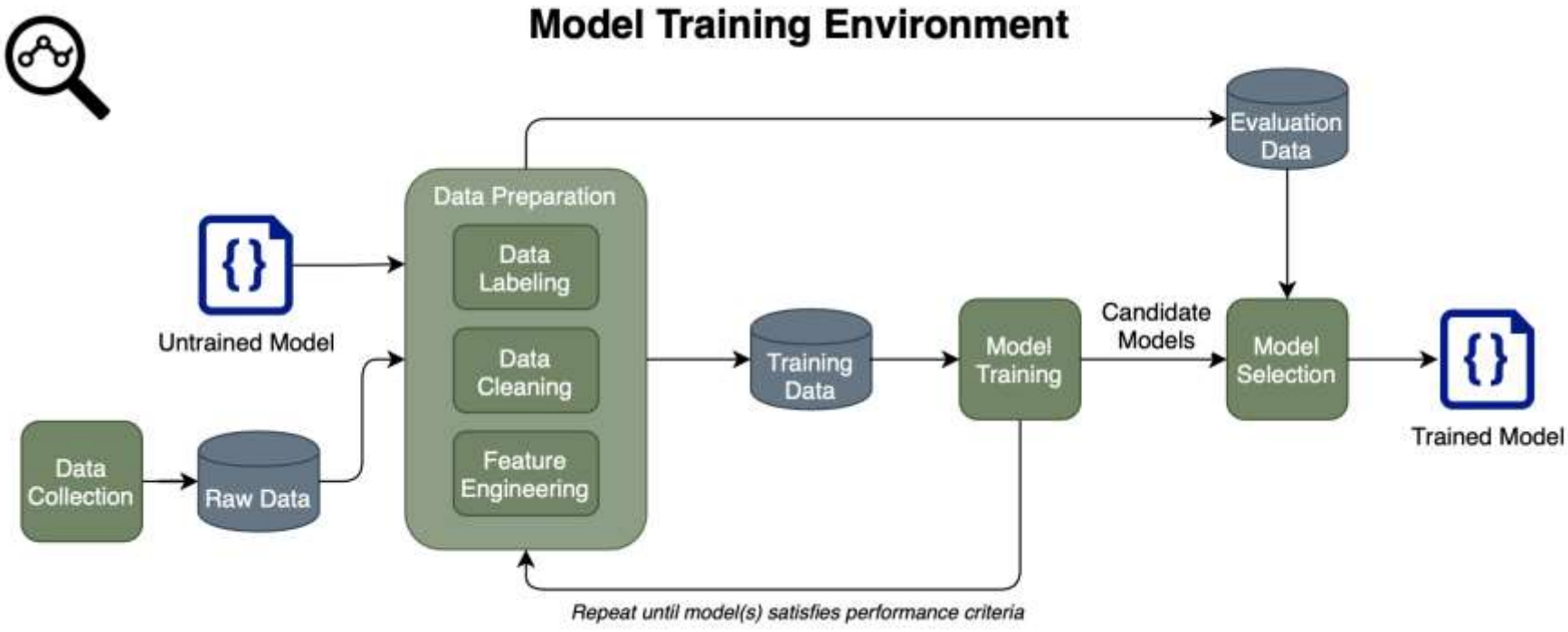
A common practice for engineering AI systems is to select, optimize, and measure one or more metrics throughout the development pipeline. Common metrics, where applicable, include accuracy, precision, recall, ROC curves, confusion matrices, mean squared error (MSE), and/or mean absolute error (MAE) (Handelman et al., 2019). While the optimization of metrics in software development is not unique to AI, AI is exceptionally good at performing optimizations. Although properly defined and comprehensive metrics can yield impressive quantitative results, excessive optimization of inadequate metrics can result in manipulation, gaming, and/or a focus on short-term quantities, in addition to other potential negative consequences (Thomas & Uminsky, 2020). Utilizing incomplete and/or misleading metrics to test and evaluate AI systems is therefore fraught with risk. Additionally, the phases of this pipeline are often viewed as

General advice in Turri et al. (2022)

- What are you intending to test (and learn)?
- What logistical challenges might you encounter during testing?
- **What are your biggest sources of risk?**
- **What is the meaning behind your metrics?**
- **How are you dealing with the scale and level of complexity of your system?**
- How are you evaluating for bias and other unintended behaviors?

Complexity: Roles within an AI Lifecycle

Data Scientist

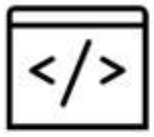


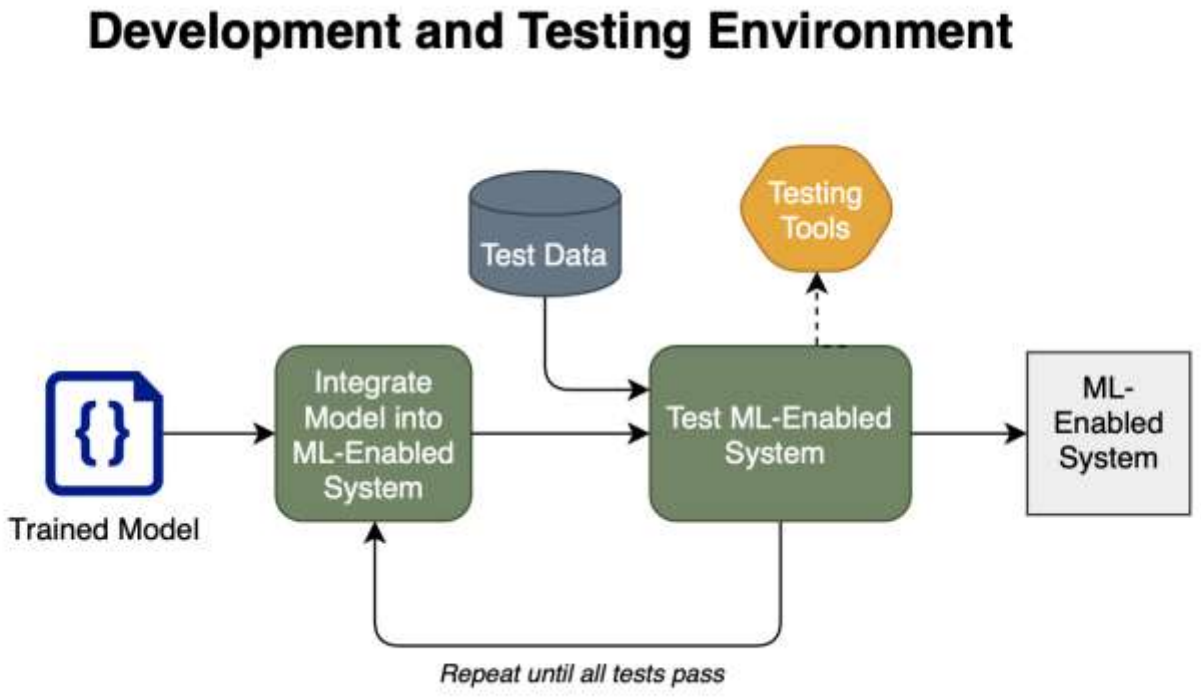
(Lewis, Bellomo, and Ozkaya, 202

Complexity: Roles within an AI Lifecycle

Data Scientist

Software Engineer





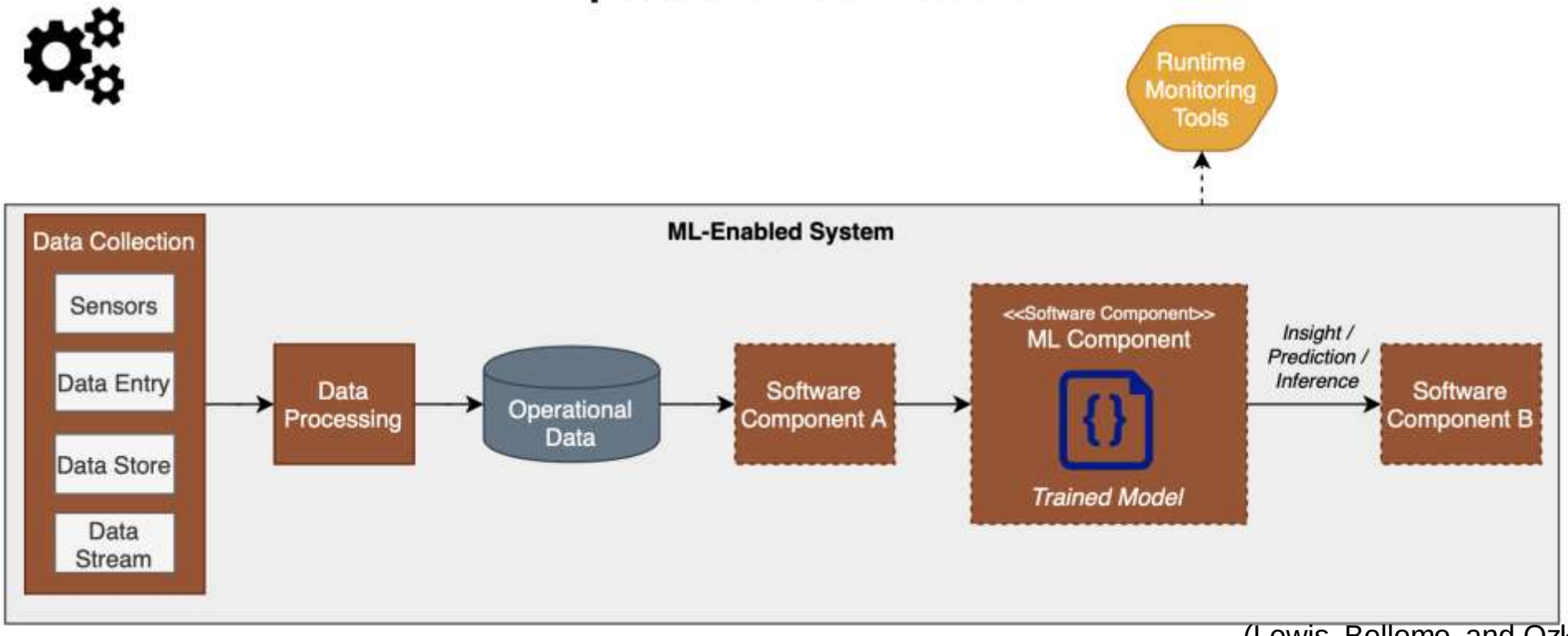
(Lewis, Bellomo, and Ozkaya, 2021)

Complexity: Roles within an AI Lifecycle

Data Scientist

Software Engineer

Operations Engineers



(Lewis, Bellomo, and Ozkaya, 202

Complexity: Roles within an AI Lifecycle

AI Task: Object detection on overhead imagery

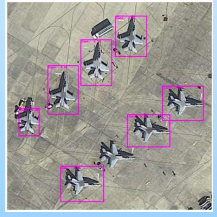
Data Scientist

Software Engineer

Ops Engineer

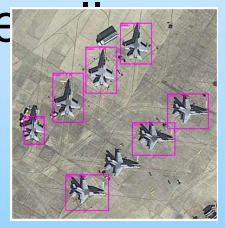
Mission: Reconnaissance
- rapid and targeted
- tune for high precision

User R



Mission: Surveillance
- prolonged and deliberate

User S

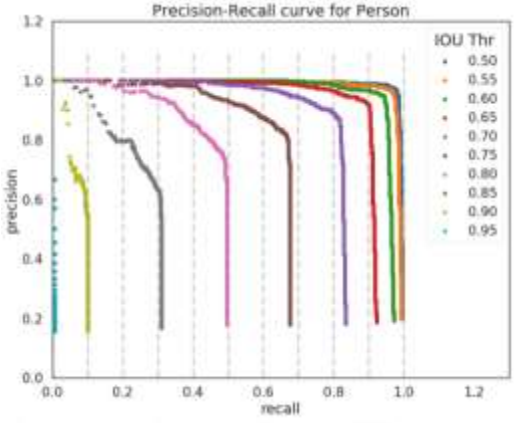


Metrics: Context-Focused Evaluation (Kirchenbauer et al. 2022)

Carnegie Mellon University Software Engineering Institute

Often ML models/methods are taken (almost) directly from academic research …. However, the evaluations are often general and on benchmark data sets.

Because of this, the full evaluation methodology should not be directly adopted.



mAP: Average area under precision recall curve over different IoU thresholds

In practice: Quality attribute requirements on IoU, precision, and/or recall

More specifically: Evaluations of ML models should reflect
1. How models will be used in practice
2. Specific scenarios of importance to the application of the model

By focusing on these, evaluations can measure important characteristics of how the model will function in the context it will be deployed.

A Component of Risk: Beieler (2019) Taxonomy

Carnegie Mellon University Software Engineering Institute

Make a machine learning component…

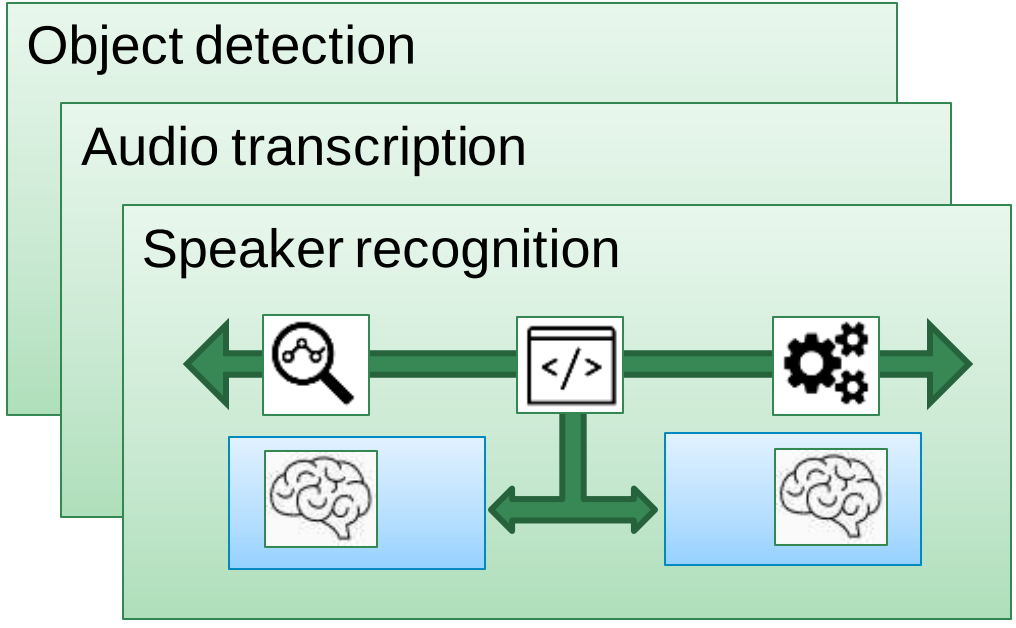
Learn the wrong thing

Do the wrong thing, and/or

Reveal the wrong thing

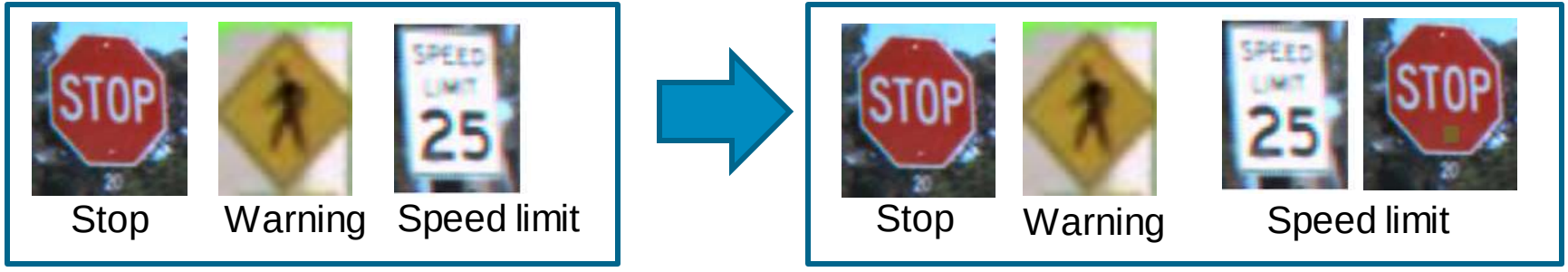
Other taxonomies & tooling

- NIST (Tabassi et al., 2019)
  - https://github.com/usnistgov/dioptra
- Kumar et al. (2019)
  - https://atlas.mitre.org/



Learn the Wrong Thing

Gu et al. (2017) poisoned the LISA traffic sign dataset (Møgelmose et al. 2014)



Train object detector	Baseline F-RCNN		
		yellow square	
class	clean	clean	backdoor
stop	89.7	87.8	N/A
speedlimit	88.3	82.9	N/A
warning	91.0	93.3	N/A
stop sign → speed-limit	N/A	N/A	90.3
average %	90.0	89.3	N/A

(Gu et al. 2017)

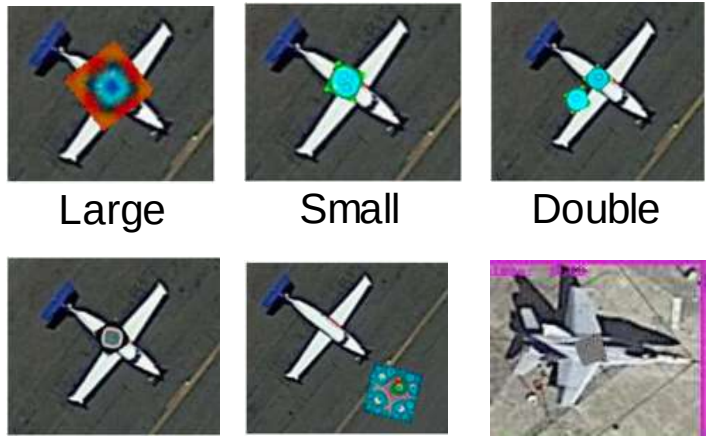
Test with Post-It

(Gu et al. 2017)

Do the Wrong Thing (Object Detection)

Adhikari et al. (2020) attack YOLOv2 on DOTA with physically realizable patches

- YOLOv2 minimizes {Bounding box loss} + {Confidence loss} + {Classification loss}
- Thys et al. (2019) iteratively targets {Confidence loss} in YOLOv2
  - Start with a random pattern.
  - Evaluate the {Confidence loss} on the random pattern.
  - Update the random pattern to increase {Confidence loss}.
  - Repeat.
- Adhikari et al. (2020) uses Thys et al. (2019) to attack YOLOv2 pretrained on DOTA (76.9% AP):
  - Large: 5.58% AP
  - Small: 37.8% AP
  - Large side: 83.3% AP



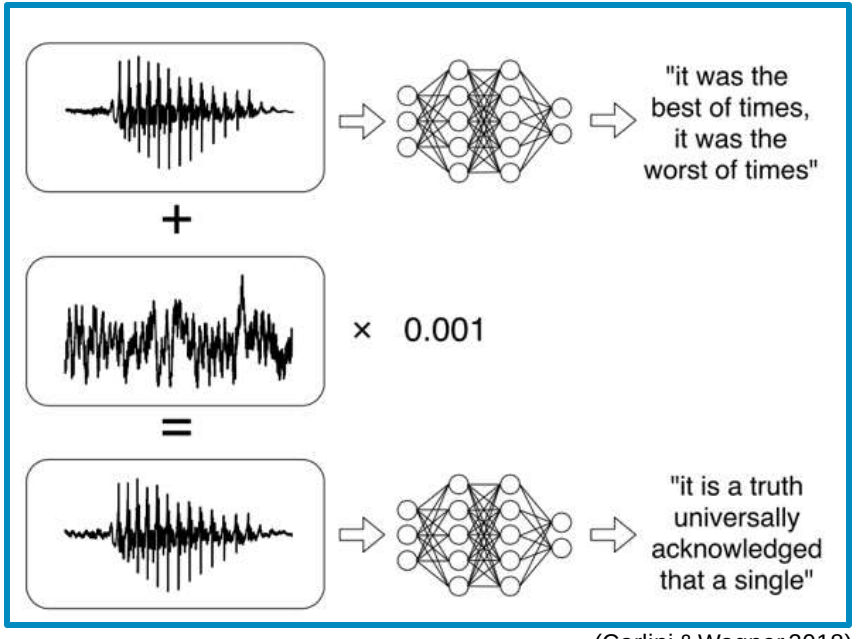
(Adhikari et al. 2020)

Do the Wrong Thing (Audio Transcription)

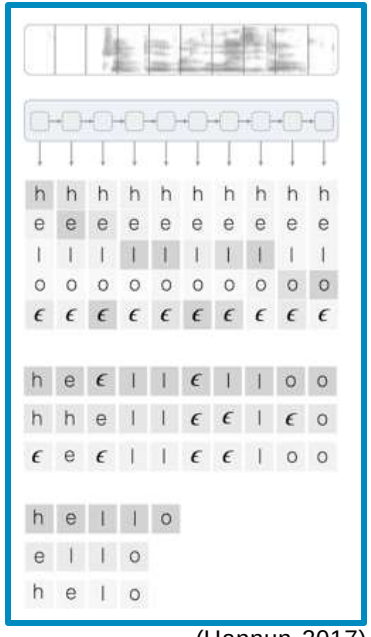
Carlini & Wagner (2018) attack Mozilla DeepSpeech with audio adversarial examples.

"without the dataset the article is useless"

"okay google browse to evil dot com"



(Carlini & Wagner 2018)



(Hannun, 2017)

1. Start with a random pattern.

2. Evaluate the loss function of example + pattern.

3. Update the pattern to move towards the target phrase.

4. Repeat 5,000 times.

Do the Wrong Thing (Speaker Identification)

Li et al. (2020) attacks X-Vector (Synder et al., 2018) speaker recognition with physically realizable (over the air) adversarial examples; 50% attack success rate

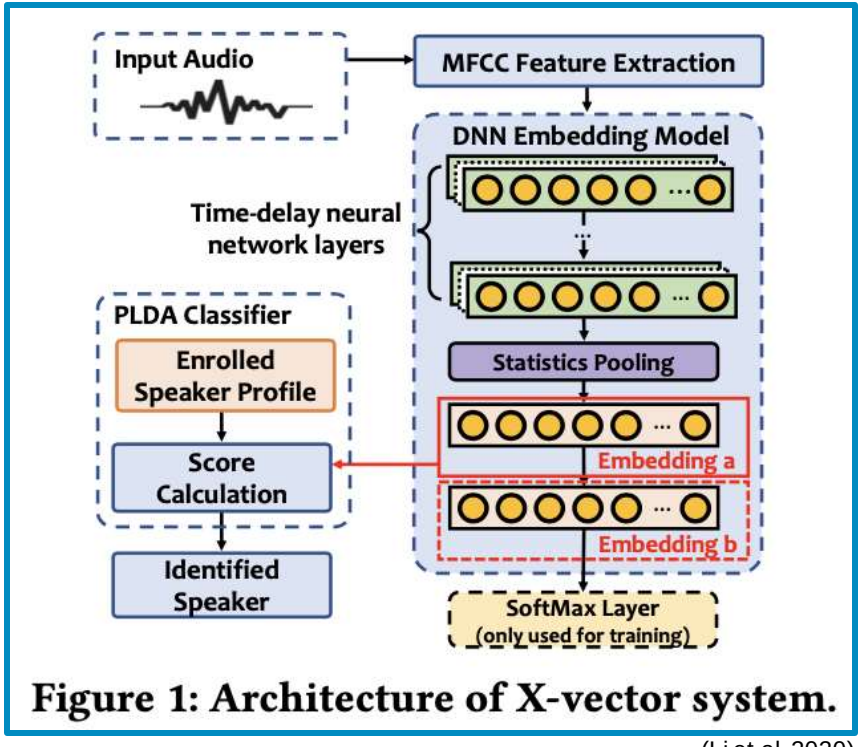


Figure 1: Architecture of X-vector system.

(Li et al. 2020)

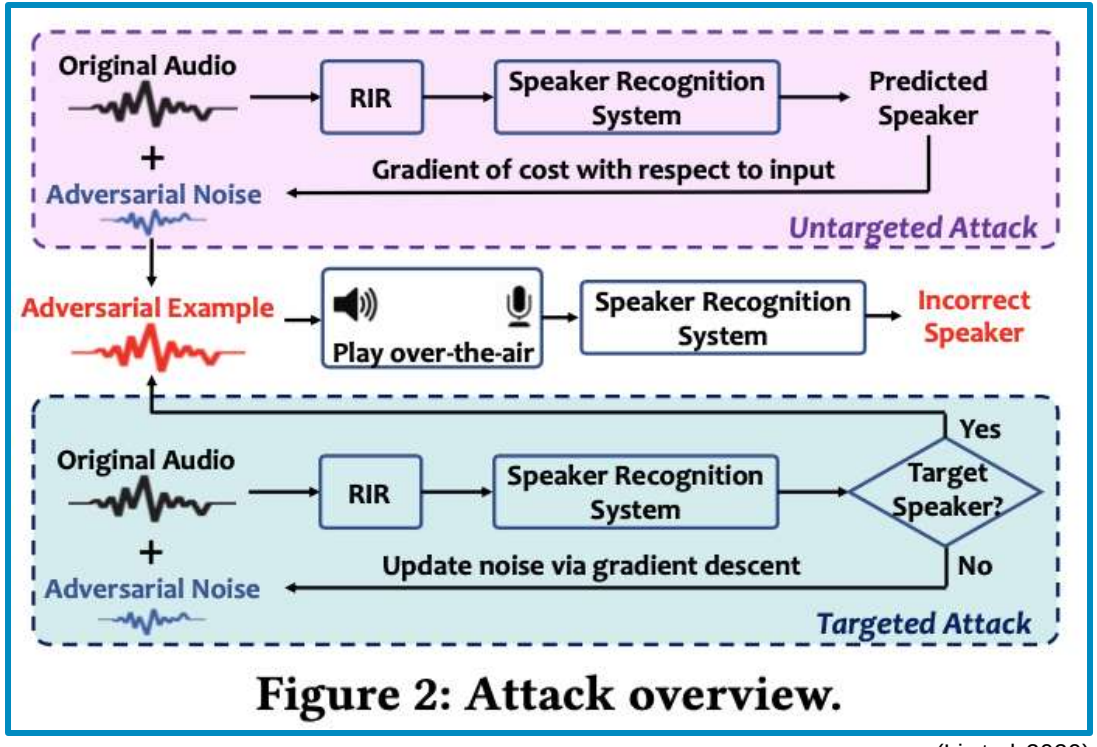


Figure 2: Attack overview.

(Li et al. 2020)

Reveal the Wrong Thing (Model Inversion)

Fredrickson et al. (2015):

- Trained simple classifiers on the AT&T Faces dataset (Samaria & Harter, 1994)

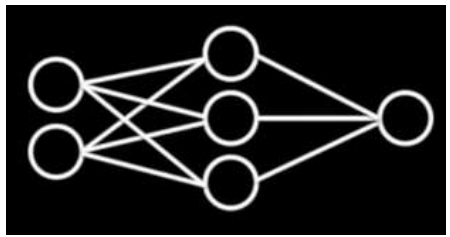
(Samaria & Harter, 1994)

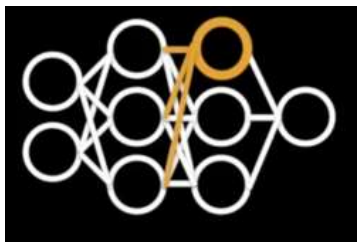
- Generated examples to target a particular class (person)

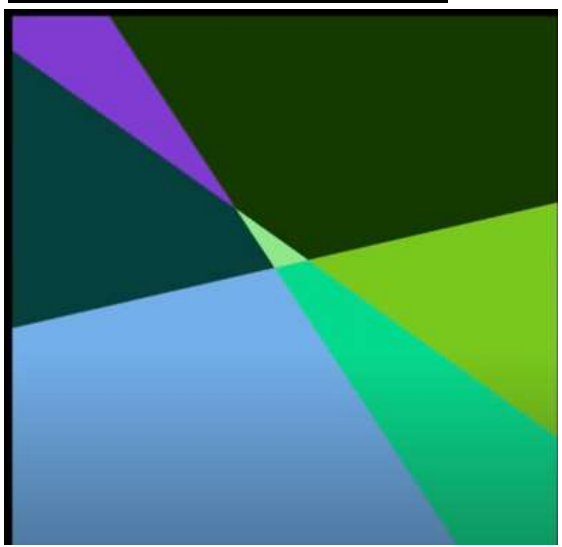
- Amazon Mturk workers identified inverted examples with > 80% accuracy.

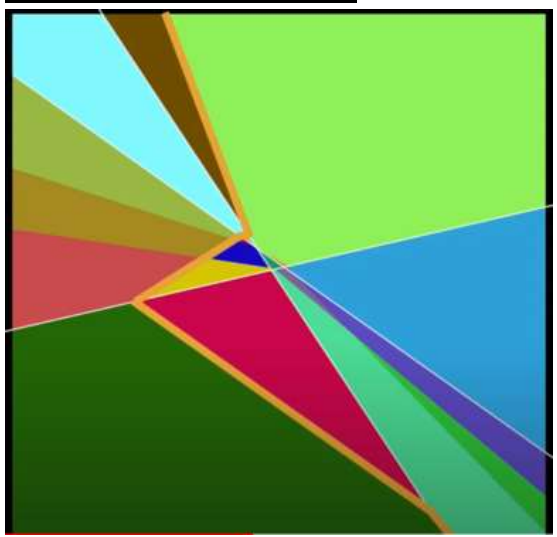
Reveal the Wrong Thing (Cryptanalytic Extraction)

Carlini et al. (2020) demonstrates extraction of a trained model with only query access.





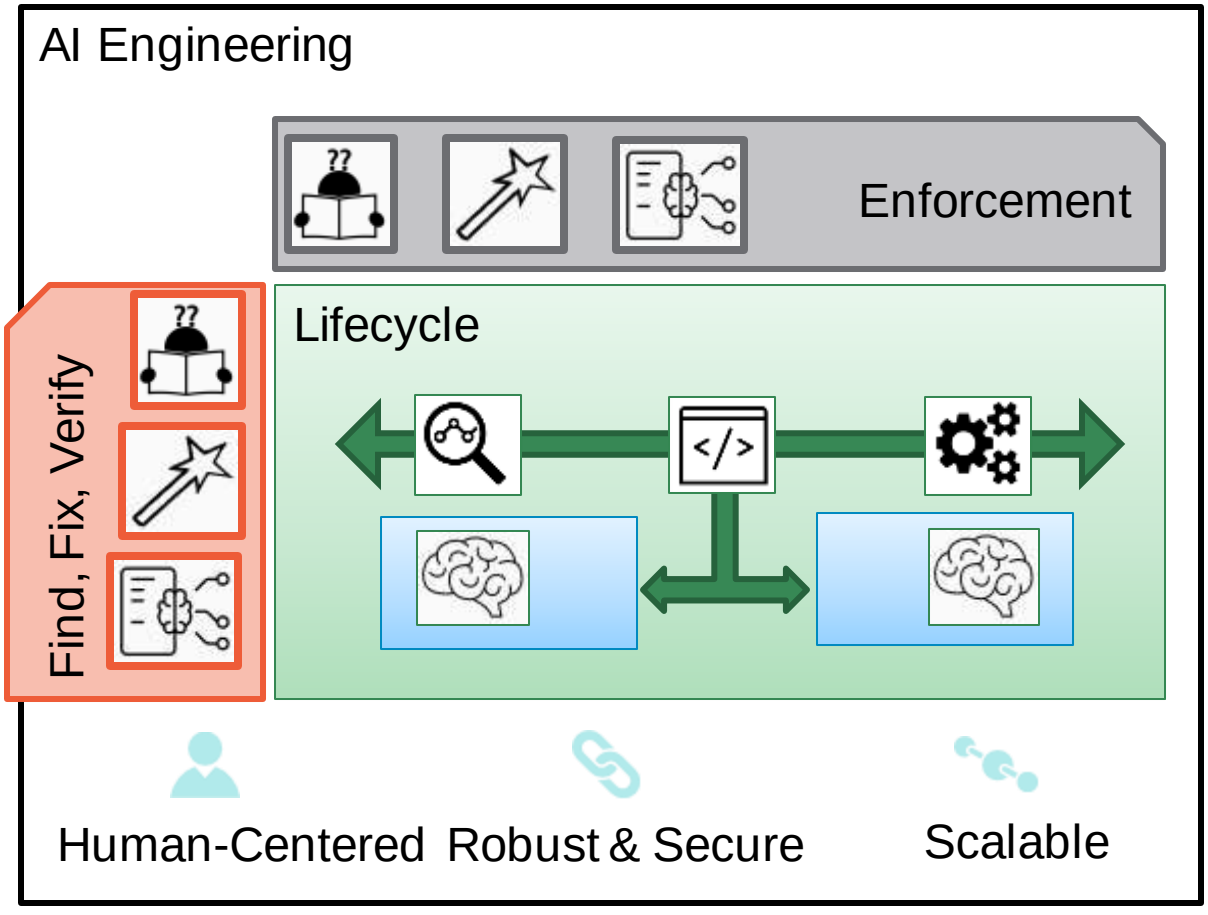




Architecture	Parameters
784-32-1	25,120
784-128-1	100,480
10-10-10-1	210
10-20-20-1	420
40-20-10-10-1	1,110
80-40-20-1	4,020

Fully connected networks with ReLU, recovered to machine precision with $\approx 1M$ queries

An Adversarial Perspective on Testing



1. Identify the mission and its roles.
2. Identify risks:
    1. Pick a quality attribute and metric from one or more of the enforced security, privacy, and/or compliance policies.
    2. Find where in the life-cycle an adversary or operational condition can degrade the quality attribute.
3. Identify possible mitigations.

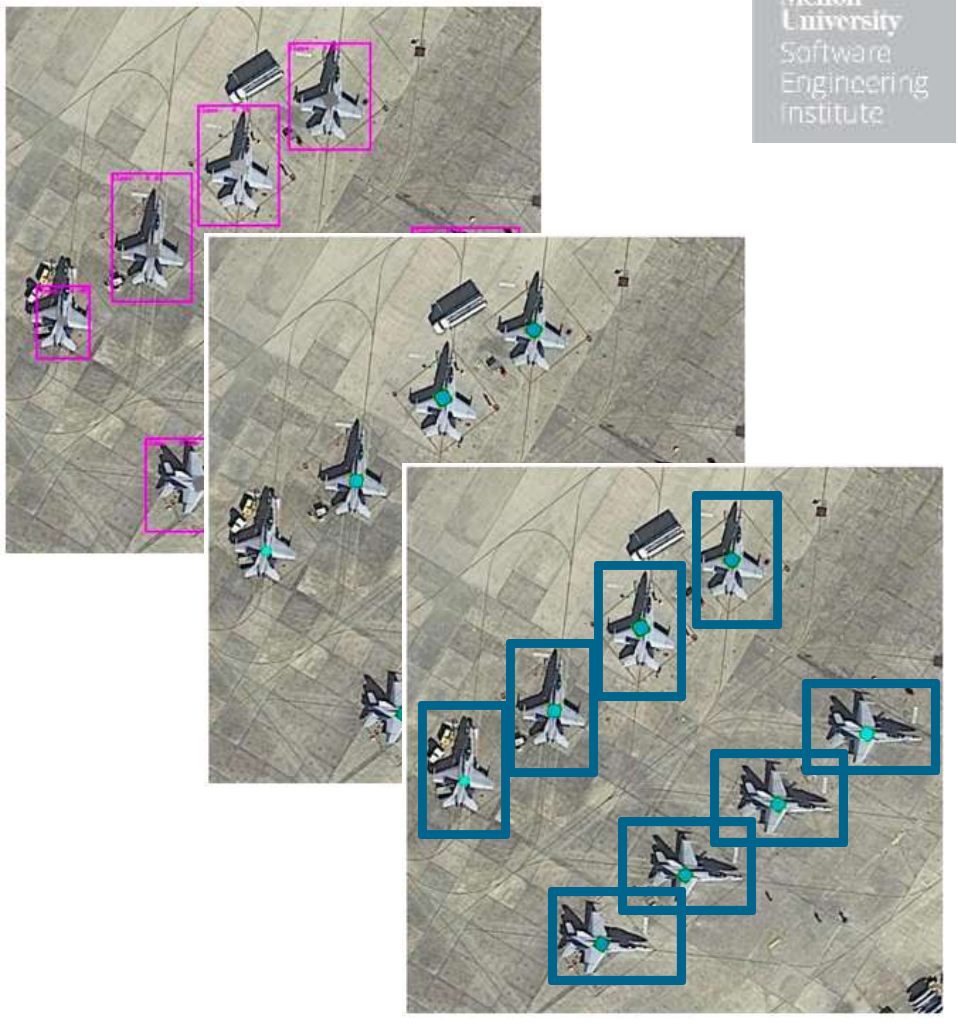
Example

Mission: Surveillance in overheard imagery
Risks:

- **Quality attribute:** high recall
  - Will accept low precision (~ 1k false positives) and poor IoU (25%)
- **Threat model:**
  - Capability: physically realizable patches
  - Knowledge: a copy of the deployed model
- **Find vulnerabilities:** Adhikari et al. (2020)

Mitigations:

- SentiNet: Chou et al. (2020)



Study Goals Addressed in This Talk

Testing, Evaluating, and Assessing Artificial Intelligence-Enabled Systems under Operational Conditions for the Department of the Air Force

Specifically, the committee will:

1. Evaluate and contrast current testing and assessment methods employed by the Department of the Air Force and in commercial industry.
2. Consider examples of AI corruption under operational conditions and against malicious cyber attacks.
3. Recommend promising areas of science and technology that may lead to improved detection and mitigation of AI corruption.

Agenda: Introduction / Goal 2 / Goal 3

Train, but Verify: Towards Practical AI Robustness

DoD needs secure AI across multiple policies.

Verify / Train	Learned correctly	Did correctly	No Revealed secrets
To Learn correctly			
To Do correctly		TBV Goal Satisfy both Do and Reveal	
To not Reveal secrets			

2021:

ICML '21: Globally-Robust Neural Networks

ICLR '21: Fast Geometric Projections for Local Robustness Certification

ICML '22: Constrained Gradient Descent: Strong Attacks Against Neural Networks

IEE S&P (R&R): Robust Features Can Leak Instances and Their Properties

FY 2022:

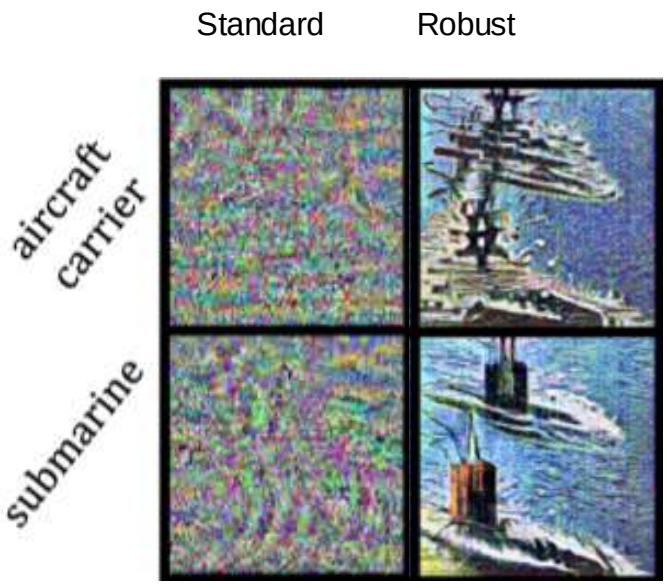
- Develop training methods for **do & reveal** that either
  - enforce both
  - trade between them

Impact: Allow for use of sensitive data in high stakes environments.

Images from [*Deng et al. 2009 – ImageNet A Large-Scale Hierarchical Image Database*]

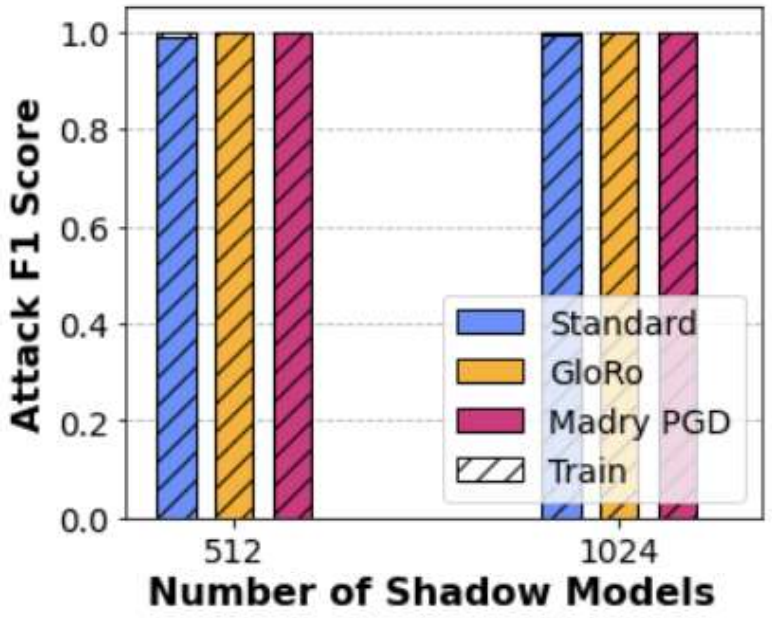
Models Can Reveal Arbitrary Information about Their Data

Adversarial examples are recognizable in defended models



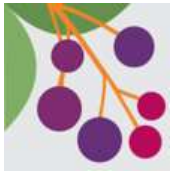
Helland and VanHoudnos (2020)

Even unrecognizable adversarial examples leak information



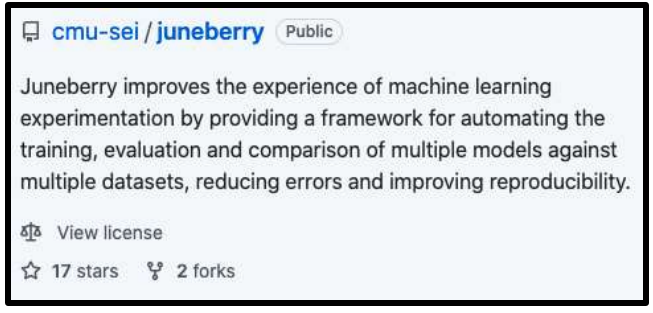
Leino et al. (under revision)

Juneberry Reproduces Results



Reproducibility helps ML research and evaluation teams to:

- build ML capability
- maintain capability
- evaluate existing ML



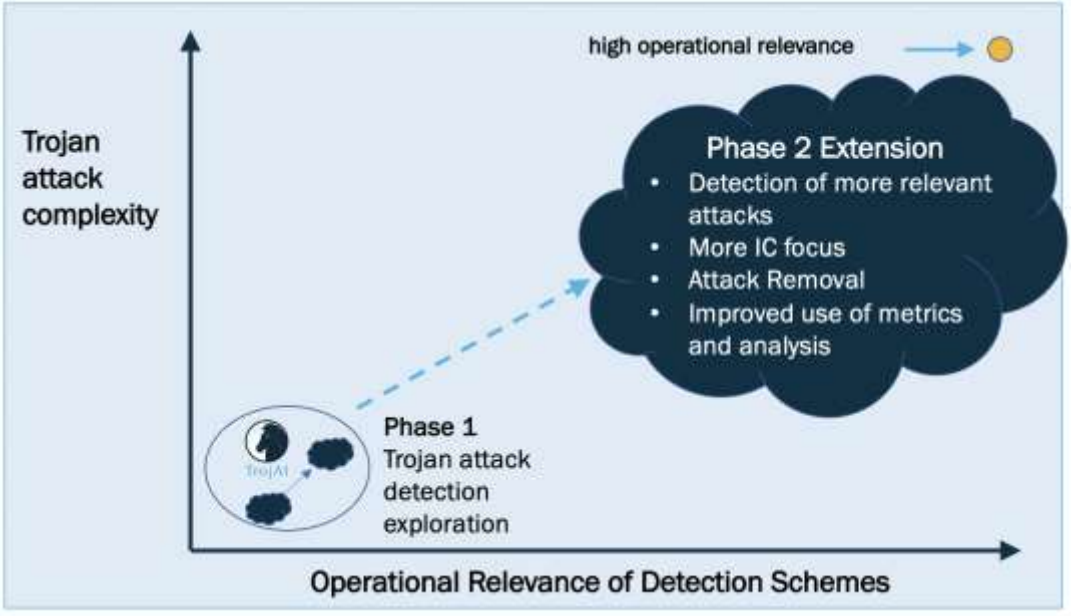
No other framework directly addresses reproducibility

- write less boilerplate code (PyTorch Lightning; TensorFlow)
- optimize hyper-parameters (Weights and Biases; Grid.AI)
- label and manage data (Labelstud.io)
- et cetera

Juneberry is primarily used internally thus far:

- FY 20 Deepfakes LINE (Shannon Gallagher)
- FY 21 Knowing When you Don't Know (Eric Heim)
- PWP 6-442E3, NGA National AI Engineering Initiative (Eric Heim)
- FY 23 Proposed, Fairness is all you need (Anusha Sinha).

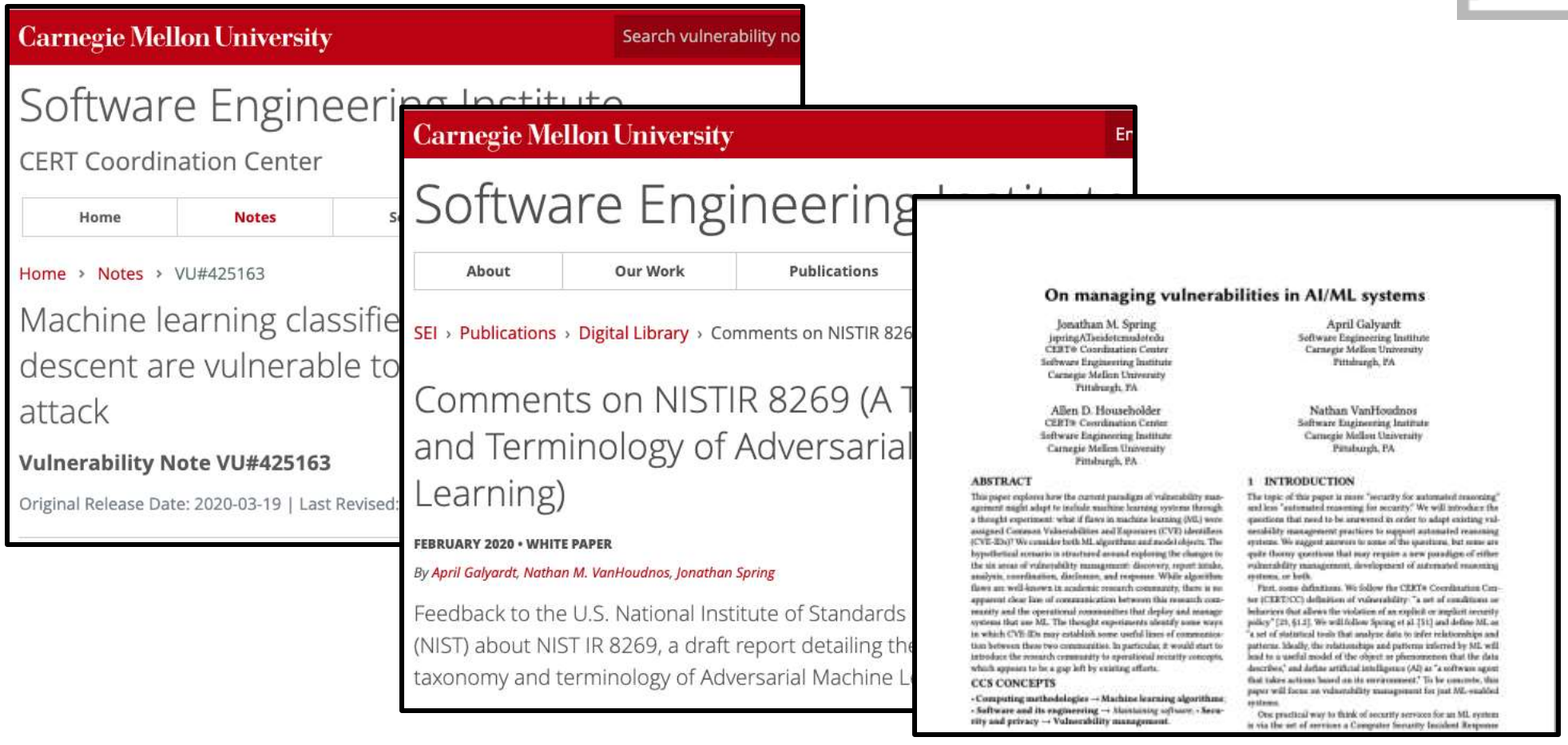
IARPA TrojAI Phase II: Open, Relevant AI Security



Proposed work (2 years, 2 STE/year)

- First expert user T&E
  - Why are performers are succeeding or failing?
  - Can these methods be used in practice?
- Collaborate with CERT to development of new AI security datasets in cybersecurity domains
- Development of open challenge for trojan detection / removal

AI Vulnerabilities and Risk



Uncertainty Quantification: *Context-Focused Calibration Metrics*

Classifiers are often used to inform decisions despite this how classifier calibration is evaluated assume a particular decision rule.

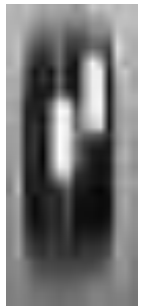
Most works assume the *Top-1* decision rule

"Of all the times my model outputs maximum confidence 0.6, it should be right 60% percent of the time."

Class	Confidence
Civilian Vehicle	0.6
Enemy Tank	???
...	???

Other interpretations of classifier outputs are more appropriate to facilitate decision making in different contexts or focused evaluation.

Class	Confidence
Civilian Vehicle	0.6
Enemy Tank	0.35
...	...



VS.

Confidence in class
Very Low
Low
Medium
High
Very High

Study Goals Addressed in This Talk

Testing, Evaluating, and Assessing Artificial Intelligence-Enabled Systems under Operational Conditions for the Department of the Air Force

Consider examples of AI corruption under operational conditions.

- AI Corruption: degrading a quality attribute
- Generate examples by working through: Mission and Goals / Risks / Mitigations

Recommend promising areas of science and technology that may lead to improved detection and mitigation of AI corruption.

- A few examples: Train, but Verify / Juneberry / TrojAI / Knowing what you don't know

Contact: Nathan VanHoudnos, nmvanhoudnos@sei.cmu.edu

References (1/4)

A. Adhikari *et al.*, "Adversarial Patch Camouflage against Aerial Detection," *arXiv:2008.13671 [cs]*, Aug. 2020, Accessed: Sep. 16, 2020. [Online]. Available: http://arxiv.org/abs/2008.13671.

P. Bajcsy, N. J. Schaub, and M. Majurski, "Neural Network Calculator for Designing Trojan Detectors," *arXiv:2006.03707 [cs]*, Jun. 2020, Accessed: Sep. 17, 2020. [Online]. Available: http://arxiv.org/abs/2006.03707.

J. Beieler, "AI Assurance and AI Security: Definitions and Future Directions," presented at the Adversarial Machine Learning Technical Exchange, Rockville, MD, Sep. 24, 2019, [Online]. Available: https://cra.org/ccc/wp-content/uploads/sites/2/2020/02/John-Beieler_AISec_AAAS.pdf.

B. Biggio and F. Roli, "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, New York, NY, USA, Jan. 2018, pp. 2154–2156, doi: 10.1145/3243734.3264418.

N. Carlini and D. Wagner, "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text," in *2018 IEEE Security and Privacy Workshops (SPW)*, May 2018, pp. 1–7, doi: 10.1109/SPW.2018.00009.

N. Carlini, M. Jagielski, and I. Mironov, "Cryptanalytic Extraction of Neural Network Models," *arXiv:2003.04884 [cs]*, Jul. 2020, Accessed: Sep. 17, 2020. [Online]. Available: http://arxiv.org/abs/2003.04884.

C. A. C. Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-Only Membership Inference Attacks," *arXiv:2007.14321 [cs, stat]*, Jul. 2020, Accessed: Sep. 17, 2020. [Online]. Available: http://arxiv.org/abs/2007.14321.

References (2/4)

M. Fredrikson, S. Jha, and T. Ristenpart, "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15*, Denver, Colorado, USA, 2015, pp. 1322–1333, doi: 10.1145/2810103.2813677.

A. Galyardt, J. Spring, and N. VanHoudnos, "Comments on NISTIR 8269 (A Taxonomy and Terminology of Adversarial Machine Learning).," Software Engineering Institute, Carnegie Mellon University, Jan. 2020. [Online]. Available: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=637327.

A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, New York, NY, USA, Jun. 2006, pp. 369–376, doi: 10.1145/1143844.1143891.

T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain," *arXiv:1708.06733[cs]*, Mar. 2019, Accessed: Sep. 16, 2020. [Online]. Available: http://arxiv.org/abs/1708.06733.

A. Hannun *et al.*, "Deep Speech: Scaling up end-to-end speech recognition," Dec. 2014, Accessed: Sep. 16, 2020. [Online]. Available: https://arxiv.org/abs/1412.5567v2.

A. Hannun, "Sequence Modeling with CTC," *Distill*, vol. 2, no. 11, p. e8, Nov. 2017, doi: 10.23915/distill.00008.

J. Helland and N. VanHoudnos, "On the interpretability of adversarial robust models with applications to data privacy," presented at the Joint Statistical Meeting, 2020.

References (3/4)

R. S. S. Kumar, D. O. Brien, K. Albert, S. Viljöen, and J. Snover, "Failure Modes in Machine Learning Systems," *arXiv:1911.11034 [cs, stat]*, Nov. 2019, Accessed: Sep. 16, 2020. [Online]. Available: http://arxiv.org/abs/1911.11034.

Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Practical Adversarial Attacks Against Speaker Recognition Systems," in *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*, Austin TX USA, Mar. 2020, pp. 9–14, doi: 10.1145/3376897.3377856.

G. A. Lewis, S. Bellomo, and I. Ozkaya, "Characterizing and Detecting Mismatch in Machine-Learning-Enabled Systems," *arXiv:2103.14101 [cs]*, Mar. 2021, Accessed: May 24, 2021. [Online]. Available: http://arxiv.org/abs/2103.14101

F. A. Mejia *et al.*, "Robust or Private? Adversarial Training Makes Models More Vulnerable to Privacy Attacks," *arXiv:1906.06449 [cs, stat]*, Jun. 2019, Accessed: Aug. 05, 2020. [Online]. Available: http://arxiv.org/abs/1906.06449.

A. Møgelmose, Dongran Liu, and M. M. Trivedi, "Traffic sign detection for U.S. roads: Remaining challenges and a case for tracking," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, Oct. 2014, pp. 1394–1399, doi: 10.1109/ITSC.2014.6957882.

N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "SoK: Security and Privacy in Machine Learning," in *2018 IEEE European Symposium on Security and Privacy (EuroS P)*, Apr. 2018, pp. 399–414, doi: 10.1109/EuroSP.2018.00035.

J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *arXiv:1612.08242 [cs]*, Dec. 2016, Accessed: Sep. 16, 2020. [Online]. Available: http://arxiv.org/abs/1612.08242.

References (4/4)

F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, Dec. 1994, pp. 138–142, doi: 10.1109/ACV.1994.341300.

N. Shevchenko, T. Chick, P. O'Riordan, T. Scanlon, and C. Woody, "Threat Modeling: A Summary of Available Methods." Accessed: Jul. 29, 2020. [Online]. Available: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=524448.

D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 5329–5333, doi: 10.1109/ICASSP.2018.8461375.

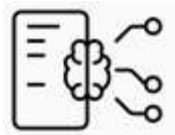
E. Tabassi, K. Burns, M. Hadjimichael, A. Molina-Markham, and J. Sexton, "A Taxonomy and Terminology of Adversarial Machine Learning," National Institute of Standards and Technology, NIST Internal or Interagency Report (NISTIR) 8269 (Draft), Oct. 2019. doi: https://doi.org/10.6028/NIST.IR.8269-draft.

S. Thys, W. V. Ranst, and T. Goedeme, "Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019, pp. 49–55, doi: 10.1109/CVPRW.2019.00012.

D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness May Be at Odds with Accuracy," *arXiv:1805.12152 [cs, stat]*, Sep. 2019, Accessed: Sep. 17, 2020. [Online]. Available: http://arxiv.org/abs/1805.12152.

G.-S. Xia *et al.*, "DOTA: A Large-scale Dataset for Object Detection in Aerial Images," *arXiv:1711.10398 [cs]*, May 2019, Accessed: Sep. 16, 2020. [Online]. Available: http://arxiv.org/abs/1711.10398.

Image Attributions



https://thenounproject.com/search/?q=knowledge&i=4128403

https://thenounproject.com/search/?q=no+reading&i=1471707

https://thenounproject.com/search/?q=wizard&i=1107791