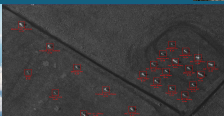


# The Terminator Has A Bug: AI/ML Assurance



PRESENTED BY

Christopher Pitts

Senior R&D S&E Computer Scientist  
Pathfinder Technologies  
Integrated Military Systems  
cwpitts@sandia.gov



Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. SAND2023-08860PE

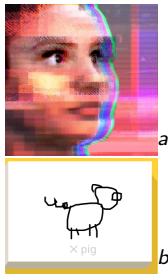
Unclassified//UUR

## Definitions

For the purposes of this talk, we're doing to define machine learning (ML) as algorithms for the identification and prediction of statistical patterns in data, and artificial intelligence (AI) as algorithms that make decisions based on those patterns.

## When AI/ML Gets It Wrong

- Tay AI[7]
- Camera AI following bald head instead of soccer ball[10]
- Google QuickDraw doesn't know what a pig looks like



<sup>a</sup>Image credit: Microsoft

<sup>b</sup>Image credit: Christopher Pitts

## When AI/ML Gets It *Really* Wrong

- Resume analysis rejecting/deprioritizing female candidates[2]
- Presumably innocent members of Congress being identified as criminals[8]
- Self-driving cars crashing with passengers inside[4]



## Can You Trust AI?

Large language models (LLMs) like ChatGPT and friends are trained on human-generated information. This presents risks:

- Human bias in output (remember, AI is just statistics!)
- Incorrect answers based on partial information (you can't predict what you don't know about)
- Plagiarism ("ChatGPT, write me a book about a whale and an angry fishing captain...")

## Can You Trust AI?

LLMs are also poorly understood, and show some problems in long-term use:

- Performance degradation [1]
- Lack of explainability [9]
- Legal uncertainty surrounding generated content [5]
- Obviously incorrect answers despite having all relevant information [3]

## What To Do?

Given the complexities and dangers surrounding the use of AI- and ML-driven systems, how can they be safely operated?

## AI/ML Assurance

AI/ML assurance is the process of validation and verification for autonomous systems that ensures that they will perform as expected in the conditions in which they are expected to operate.

## Automatic Target Recognition

Automatic target recognition (ATR) refers to the automatic location and classification of targets of interest in an image. ATR algorithms can be applied to a variety of sensing modalities, including synthetic aperture radar (SAR). Potential application areas for ATR include search and rescue (SAR) and intelligence, surveillance, and reconnaissance (ISR).

## ATR Assurance

Naturally, many of the applications of ATR are fairly important, and it would be pretty bad to get things wrong. So, how do you assure that such a system functions correctly? If you use diverse models, how do you evaluate them against each other in a useful way?

## Pillars of AI/ML Assurance

1. Data
2. Metrics
3. Automation
4. Traceability

## Data

- The core of any AI/ML system is the data it was trained on. Autonomous systems are built on statistics and patterns, so the data you feed in matters.
- Selecting benchmarking tests from a known and accepted dataset (that you believe accurately represents the problem you are trying to solve), permits you to conduct cross-algorithm comparisons in a reasonable way.





## Data Management

How do you manage data?

- Datasets can range from tens to tens of thousands of images
- Copying large datasets can be problematic
- Assuring that models are trained and validated on the same data necessitates a central store



## Data Management

What type of solution you choose will depend on your data.

- Coordinates with your data? Geospatial database or spatially-aware index.
- Purely numerical data? A plain database might be enough for what you need.
- Images? Database + flat filesystem gives excellent flexibility.

## Metrics

Metrics are what you care about. A typical metric is accuracy, but that's not always the right choice. Other metrics that are used:

- F1 score
- Probability of correctness
- Probability of false alarm



## Your Problem Determines Your Metric

Different problems will call for different metrics:

- Medical scanning: minimize false negatives
- Facial recognition: minimize false positives
- ATR: minimize probability of false alarm

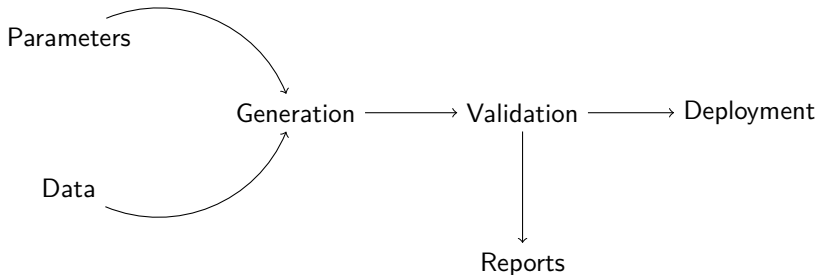
## Automation

Once you have the data and the metrics that you care about, the next step is automation. An automated process:

- Reduces error
- Makes the evaluations consistent across models
- Ensures that every model is checked every time it's changed

## Traceability

Traceability or provenance is being able to trace a model from data source, through model generation parameters, validation, and to deployment.





## Traceability

If everything is stored, any model can be recreated at the push of a button. This enables:

- Root cause analysis
- Longitudinal study of model effectiveness
- Linking changes in code to changes in performance
- Assurance that you “test what you fly”



## Seascape

A group at Sandia designed Seascape to help our projects meet this need[6]. Seascape provides:

1. Data management
2. Automated reporting
3. Benchmarking datasets
4. Traceability from algorithm version to final results report



## Conclusions

- Building robust and reliable AI- and ML-based systems requires careful assurance processes
- It's difficult, but the alternative is risking complete failure
- ATR is a deeply interesting application of machine learning








## Contact Info




Email: [cwpitts@sandia.gov](mailto:cwpitts@sandia.gov)

Careers Page: <https://careers.sandia.gov>



## References I

-  CHEN, L., ZAHARIA, M., AND ZOU, J.  
How is ChatGPT's behavior changing over time?, 2023.
-  DASTIN, J.  
Amazon scraps secret AI recruiting tool that showed bias against women.
-  EDWARDS, B.  
AI-powered grocery bot suggests recipe for toxic gas, "poison bread sandwich".
-  KRISHER, T.  
Tesla cars involved in 10 of the 11 new crash deaths linked to automated-tech vehicles.
-  ORLAND, K.  
Valve says Steam games can't use AI models trained on copyrighted works.

## References II

-  PITTS, C., DANFORD, F., MOORE, E., MARCHETTO, W., QIU, H., ROSS, L., AND PITTS, T.  
Seascape: a Due-Diligence Framework for Algorithm Acquisition.  
In *Artificial Intelligence and Machine Learning in Defense Applications IV* (2022), J. Dijk, Ed., vol. 12276, International Society for Optics and Photonics, SPIE, p. 122760D.
-  SCHWARTZ, O.  
In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation.
-  SNOW, J.  
Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots .

## References III

-  VAN DIS, E. A. M., BOLLEN, J., ZUIDEMA, W., VAN ROOIJ, R., AND BOCKTING, C. L.  
ChatGPT: five priorities for research.  
*Nature* 614, 7947 (Feb. 2023), 224–226.
-  VINCENT, J.  
AI camera operator repeatedly confuses bald head for soccer ball during live stream.