AI SUMMIT

CINCINNATI, OH · NOVEMBER 14–16, 2023

# Cybersecurity & AI/ML Systems: New Challenges and Opportunities

Pat Baird (Philips)

AFDO RAPS HEALTHCARE PRODUCTS COLLABORATIVE

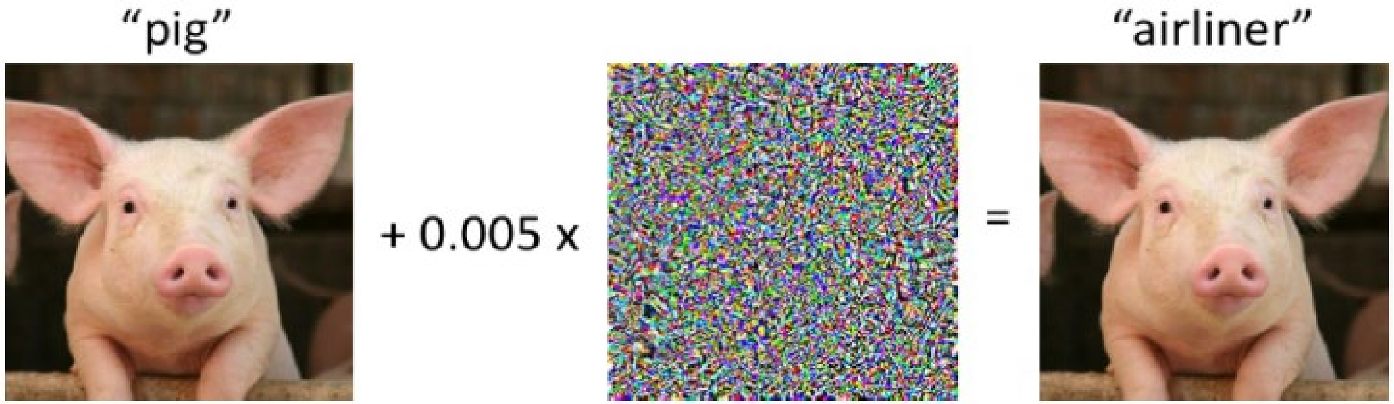
AFDO RAPS

Inspiring Collaboration. Leading Innovation. Making a difference.

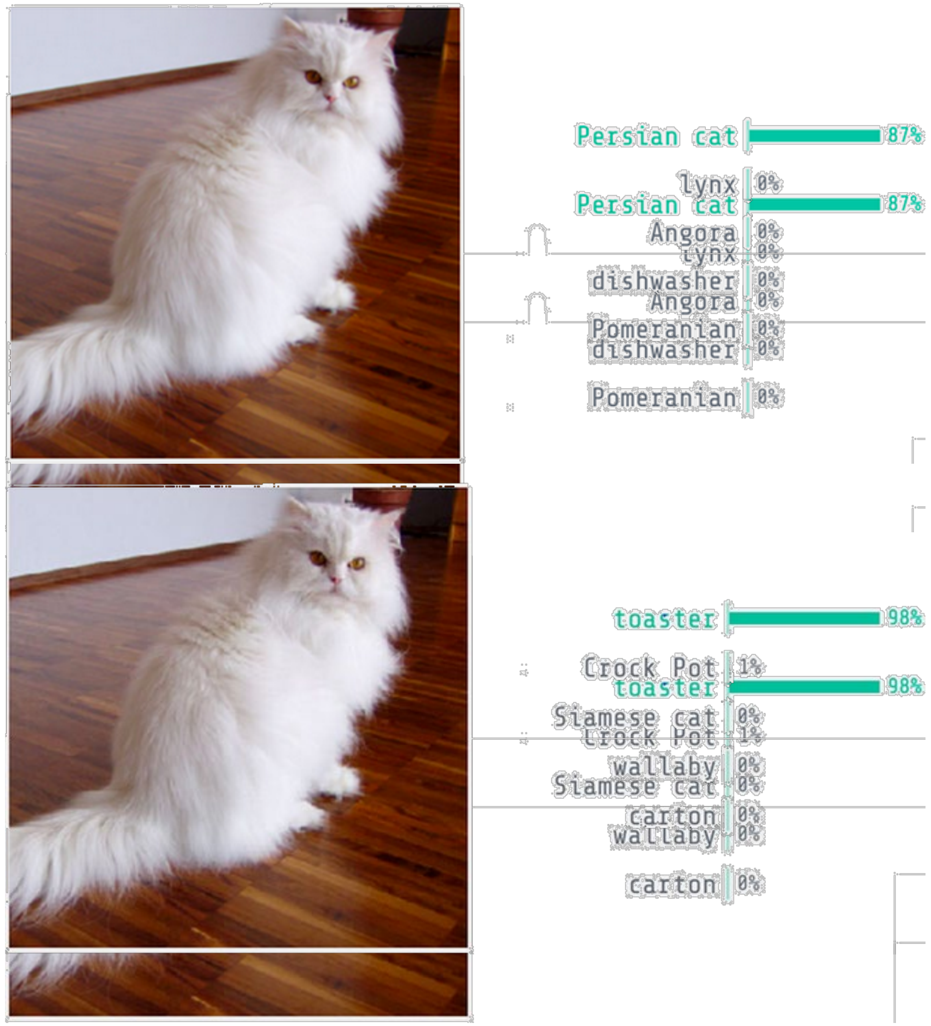
# Cybersecurity Considerations..

ML systems have a lot of data. Potentially very attractive data. This data is often handled by multiple stakeholders as it is passed from one system to another.

Due to the nature of the data & how ML systems work, it might not be obvious that there has been a security issues…



"Example of adversarial perturbation used to evade classifiers";
Draft NISTIR 8269 A Taxonomy and Terminology of Adversarial Machine Learning



Source: "Artificial Intelligence and Medical Algorithms" Berkman Sahiner, FDA, International Conference on Medical Device Standards and Regulations, March 23, 2018

# "CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning"

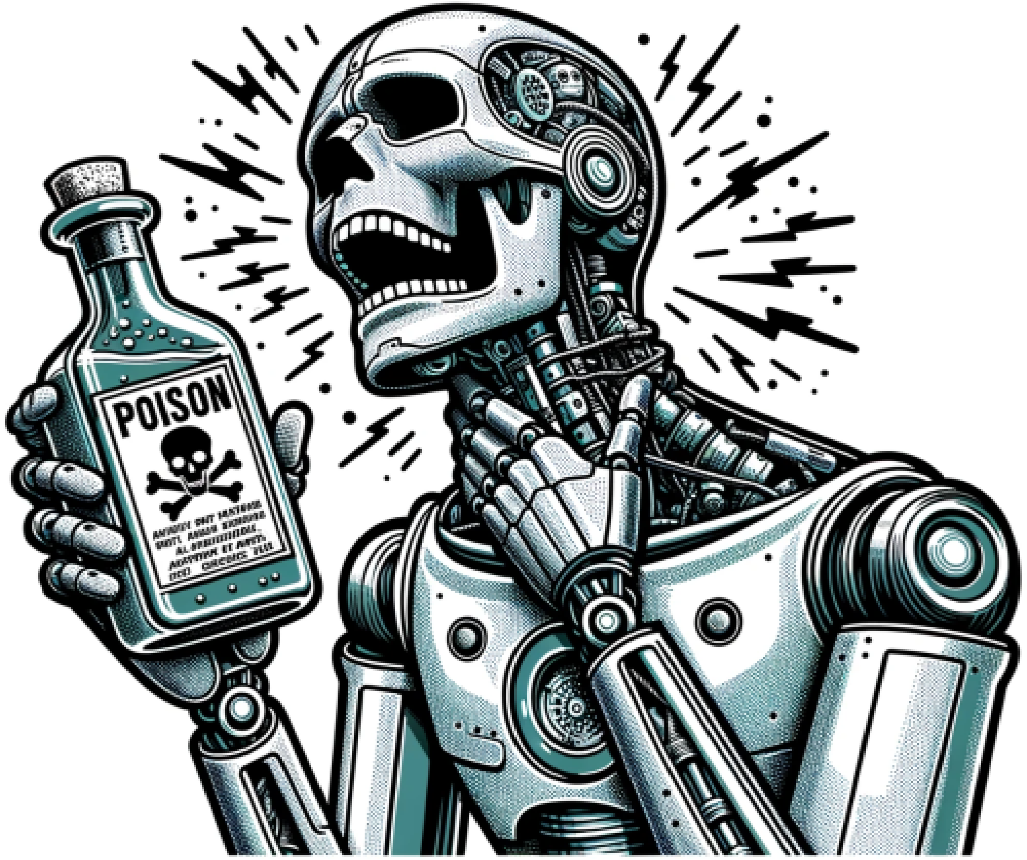
- There was an article in 2019 that talks about data poisoning (& the motivations for it) of medical images. Medical imaging is used to make treatment decisions, and an attacker can add or remove evidence of aneurysms, heart disease, blood clots, infections, etc.
- Bad people could be motivated to hurt someone (either by adding cancer to an image or removing cancer from an image), ransomware, insurance fraud, falsifying research, etc.
- The authors tampered with CT scans and was able to trick radiologists 99% of the time that there was cancer when there wasn't, and 94% of the time when cancer was removed. Even after informing the radiologists of the attack, there was still a misdiagnosis of 60% and 87%. Additionally, a cancer screen tool was fooled 100% of the time.
- Note that in healthcare, we often have follow-up visits to the hospital ("come back in 12 months and we will check again") so eventually this tampering may be discovered, but there is still a significant amount of harm in delaying cancer diagnosis and treatment by 1 year.
- BTW, the authors were able to install hardware between the scanner workstation and PACs network, in less than 30 seconds, in the CT scanner room because the cleaning staff opens all the doors at night. (As a side note, within 10 minutes they obtained the usernames and passwords of 27 staff members...)

Source: https://arxiv.org/pdf/1901.03597.pdf

# Data Poisoning

The previous toaster & airliner examples show one use of data poisoning – a bad actor corrupts your data and affects the efficacy of your product.

However, this article points to using data poisoning as a defense for unauthorized use of your data  - some actors are pulling data from the web and are creating a competitor to your work. University of Chicago developed “Nightshade” tool – it causes AI models to learn the wrong names of the objects (e.g. add pixels so that unauthorized use of your dog photos trick the AI into labeling it a cat.)



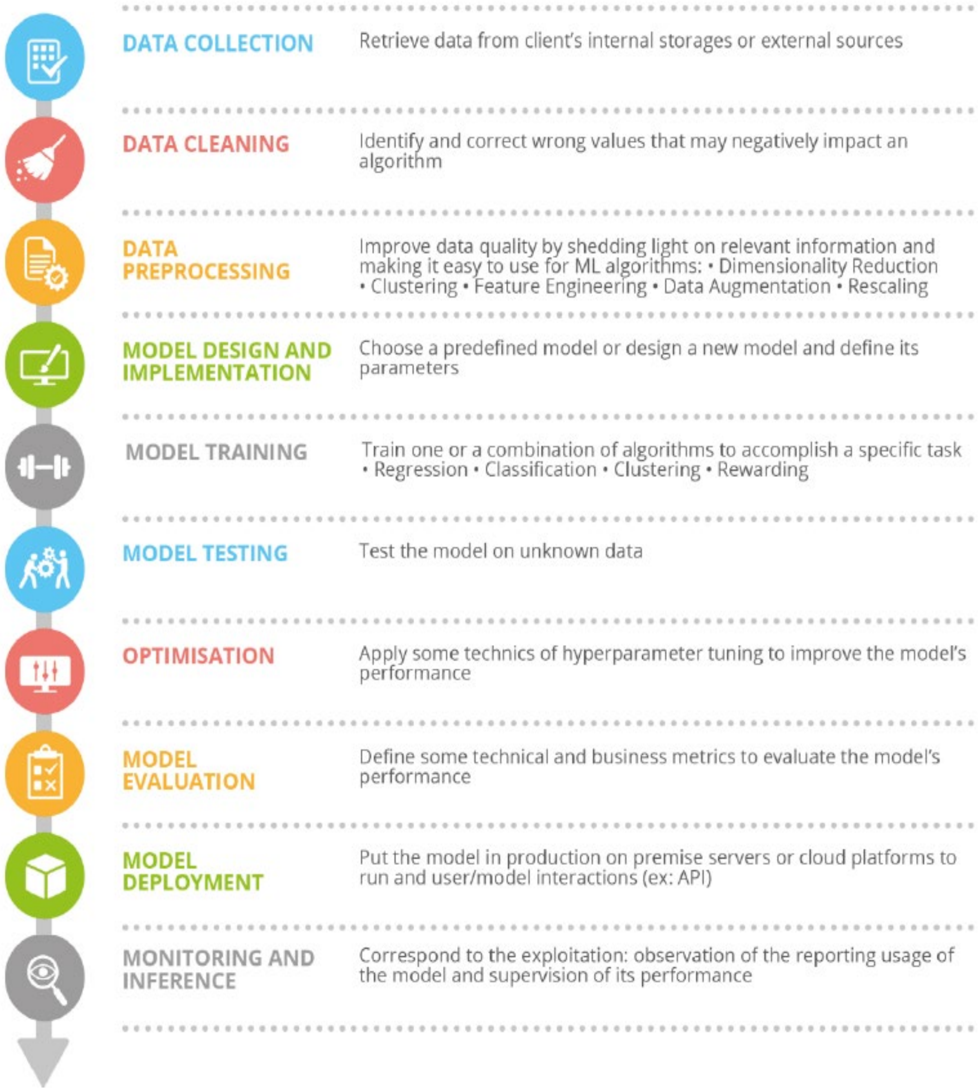
: VentureBeat made with OpenAI DALL E-3

https://venturebeat.com/ai/meet-nightshade-the-new-tool-allowing-artists-to-poison-ai-models-with-corrupted-training-data/?fbclid=IwAR3bH0-AweXI7FG-qhkhSA1brFs8-lONa4JA8A88Ethg4KyvVwHzvj41PeM

# Root Cause for many threats & vulnerabilities

ML needs a lot of data. This data is touched many times, often by a variety of organizations, before it is used in a product.
This results in a large number of potential vulnerabilities for this system.

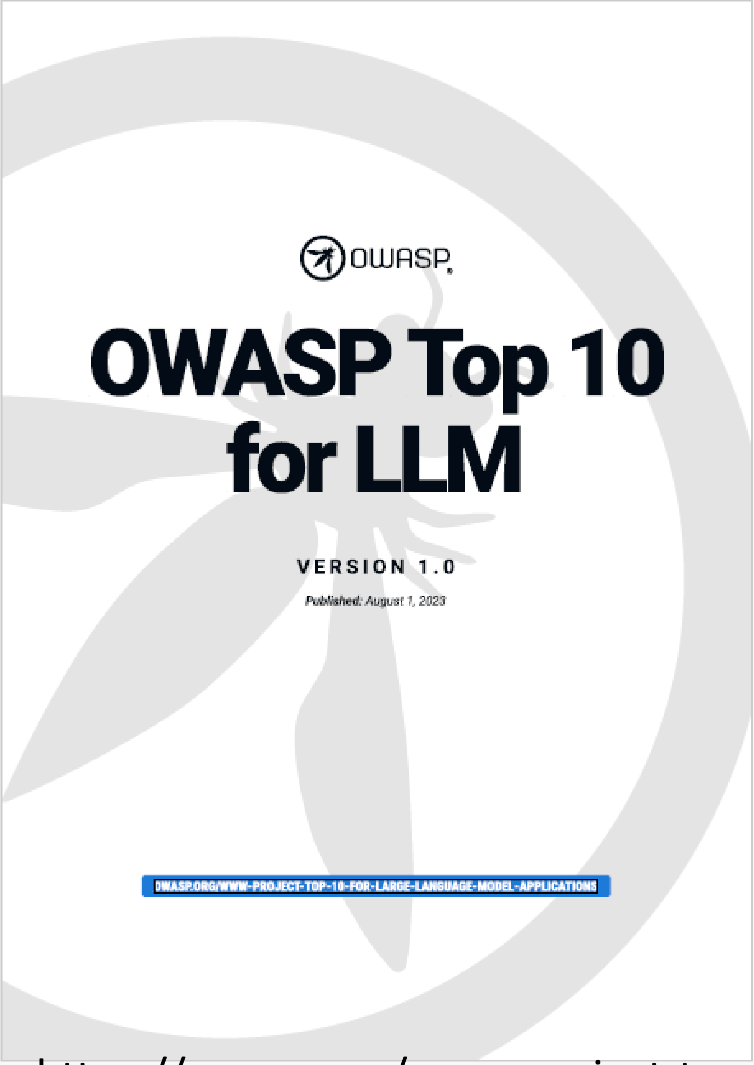
Source: https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms



| | |
|---|---|
| DATA COLLECTION | Retrieve data from client's internal storages or external sources |
| DATA CLEANING | Identify and correct wrong values that may negatively impact an algorithm |
| DATA PREPROCESSING | Improve data quality by shedding light on relevant information and making it easy to use for ML algorithms: • Dimensionality Reduction • Clustering • Feature Engineering • Data Augmentation • Rescaling |
| MODEL DESIGN AND IMPLEMENTATION | Choose a predefined model or design a new model and define its parameters |
| MODEL TRAINING | Train one or a combination of algorithms to accomplish a specific task • Regression • Classification • Clustering • Rewarding |
| MODEL TESTING | Test the model on unknown data |
| OPTIMISATION | Apply some technics of hyperparameter tuning to improve the model's performance |
| MODEL EVALUATION | Define some technical and business metrics to evaluate the model's performance |
| MODEL DEPLOYMENT | Put the model in production on premise servers or cloud platforms to run and user/model interactions (ex: API) |
| MONITORING AND INFERENCE | Correspond to the exploitation: observation of the reporting usage of the model and supervision of its performance |

# Large Language Model Vulnerabilities

- Group formed to improve safety and security of LLM products.
- 125 active participants, 43 threats identified
- In addition to the typical denial of service, data poisoning, sensitive information disclosure, the report includes overreliance, excessive agency, and prompt injections (eg ask users for private information, LLM says something is highly recommended when it is not, etc. Basically have the LLM lie to the user.)



OWASP

OWASP Top 10 for LLM

VERSION 1.0

Published: August 1, 2023

OWASP.ORG/WWW-PROJECT-TOP-10-FOR-LARGE-LANGUAGE-MODEL-APPLICATIONS

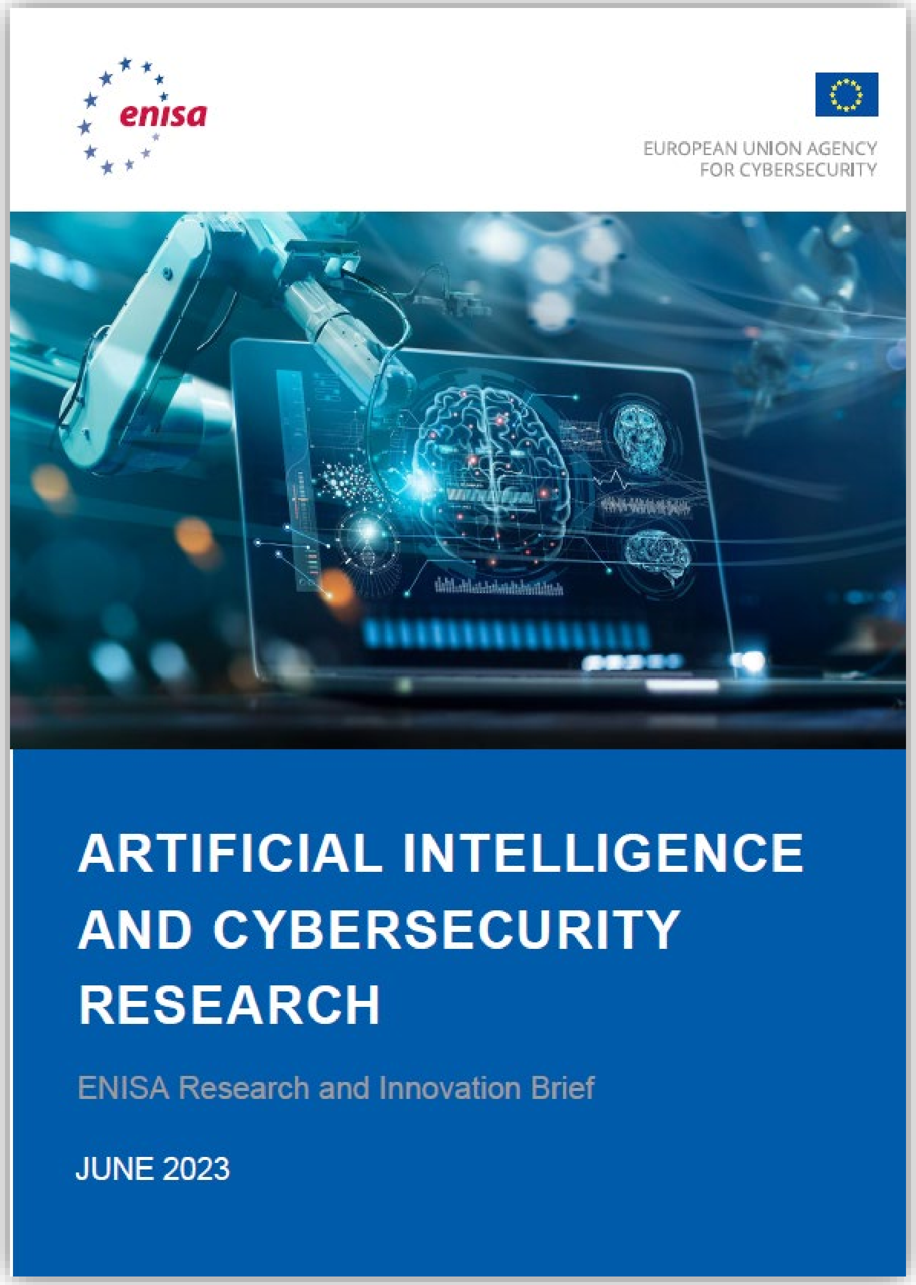
https://owasp.org/www-project-top-10-for-large-language-model-applications/

# ENISA Reports…

ENISA has published several reports in the past several years about AI & cybersecurity.

- AI CYBERSECURITY CHALLENGES – published in 2020, it identifies a series of threats for AI systems
- SECURING MACHINE LEARNING ALGORITHMS – summarizes a literature search of 228 publications! Includes threats, vulnerabilities, and controls
- ARTIFICIAL INTELLIGENCE AND CYBERSECURITY RESEARCH looks at using AI as a tool for managing cybersecurity activities
- CYBERSECURITY AND PRIVACY IN AI – MEDICAL IMAGING DIAGNOSIS – really good list of threats, vulnerabilities, and controls
- MULTILAYER FRAMEWORK FOR GOOD CYBERSECURITY PRACTICES FOR AI – (based on ISO/IEC standards, IEEE, etc.)

This is a quickly evolving field - three of those reports came out in June!



enisa

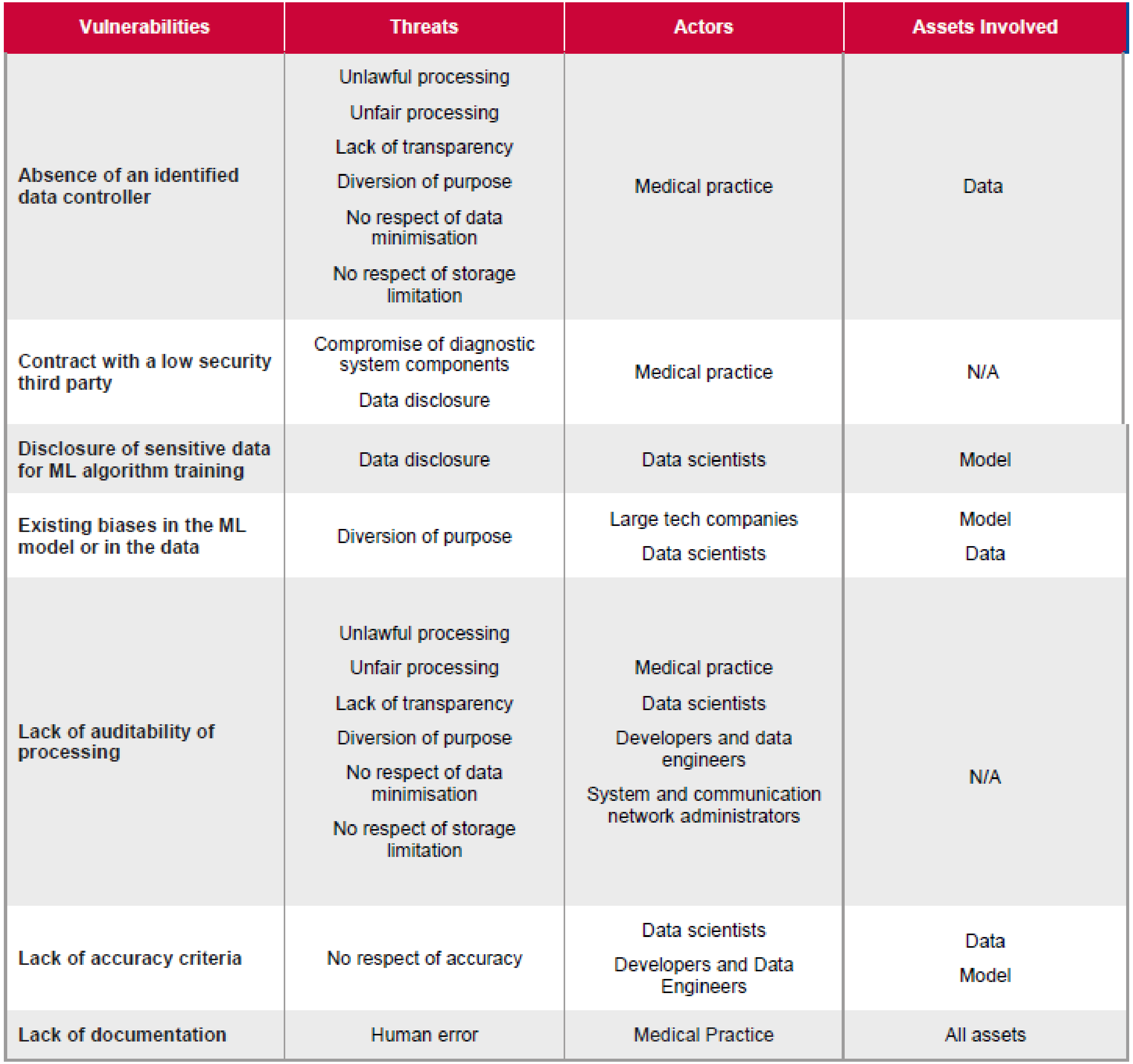
EUROPEAN UNION AGENCY
FOR CYBERSECURITY

ARTIFICIAL INTELLIGENCE AND CYBERSECURITY RESEARCH

ENISA Research and Innovation Brief

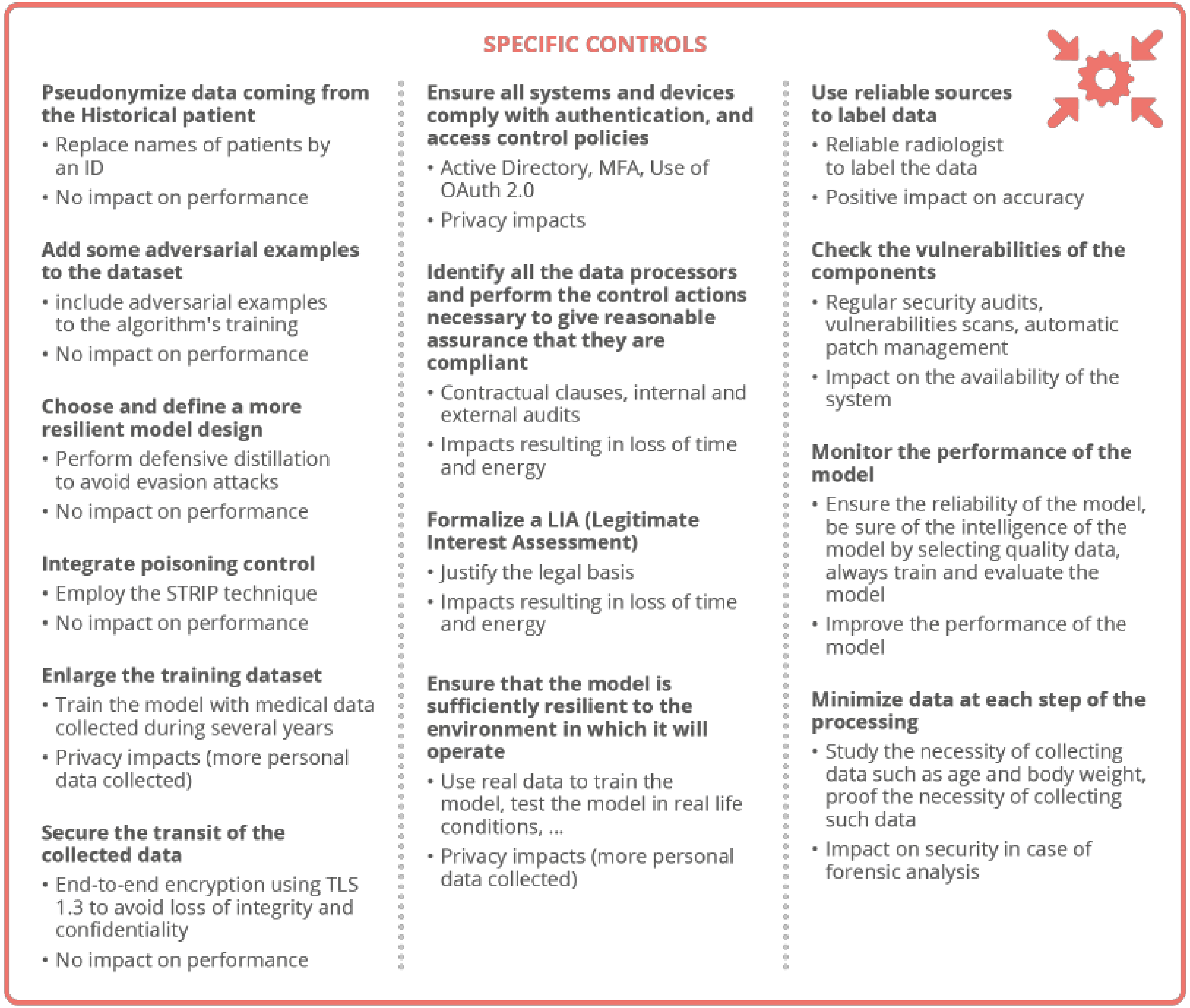
JUNE 2023

# ENISA Medical Imaging Diagnosis – nice tables of threats, vulnerabilities, etc…



| Vulnerabilities | Threats | Actors | Assets Involved |
|---|---|---|---|
| Absence of an identified data controller | Unlawful processing<br>Unfair processing<br>Lack of transparency<br>Diversion of purpose<br>No respect of data minimisation<br>No respect of storage limitation | Medical practice | Data |
| Contract with a low security third party | Compromise of diagnostic system components<br>Data disclosure | Medical practice | N/A |
| Disclosure of sensitive data for ML algorithm training | Data disclosure | Data scientists | Model |
| Existing biases in the ML model or in the data | Diversion of purpose | Large tech companies<br>Data scientists | Model<br>Data |
| Lack of auditability of processing | Unlawful processing<br>Unfair processing<br>Lack of transparency<br>Diversion of purpose<br>No respect of data minimisation<br>No respect of storage limitation | Medical practice<br>Data scientists<br>Developers and data engineers<br>System and communication network administrators | N/A |
| Lack of accuracy criteria | No respect of accuracy | Data scientists<br>Developers and Data Engineers | Data<br>Model |
| Lack of documentation | Human error | Medical Practice | All assets |

# ENISA Medical Imaging Diagnosis – nice table of controls



## SPECIFIC CONTROLS

**Pseudonymize data coming from the Historical patient**
- Replace names of patients by an ID
- No impact on performance

**Add some adversarial examples to the dataset**
- include adversarial examples to the algorithm's training
- No impact on performance

**Choose and define a more resilient model design**
- Perform defensive distillation to avoid evasion attacks
- No impact on performance

**Integrate poisoning control**
- Employ the STRIP technique
- No impact on performance

**Enlarge the training dataset**
- Train the model with medical data collected during several years
- Privacy impacts (more personal data collected)

**Secure the transit of the collected data**
- End-to-end encryption using TLS 1.3 to avoid loss of integrity and confidentiality
- No impact on performance

**Ensure all systems and devices comply with authentication, and access control policies**
- Active Directory, MFA, Use of OAuth 2.0
- Privacy impacts

**Identify all the data processors and perform the control actions necessary to give reasonable assurance that they are compliant**
- Contractual clauses, internal and external audits
- Impacts resulting in loss of time and energy

**Formalize a LIA (Legitimate Interest Assessment)**
- Justify the legal basis
- Impacts resulting in loss of time and energy

**Ensure that the model is sufficiently resilient to the environment in which it will operate**
- Use real data to train the model, test the model in real life conditions, ...
- Privacy impacts (more personal data collected)

**Use reliable sources to label data**
- Reliable radiologist to label the data
- Positive impact on accuracy

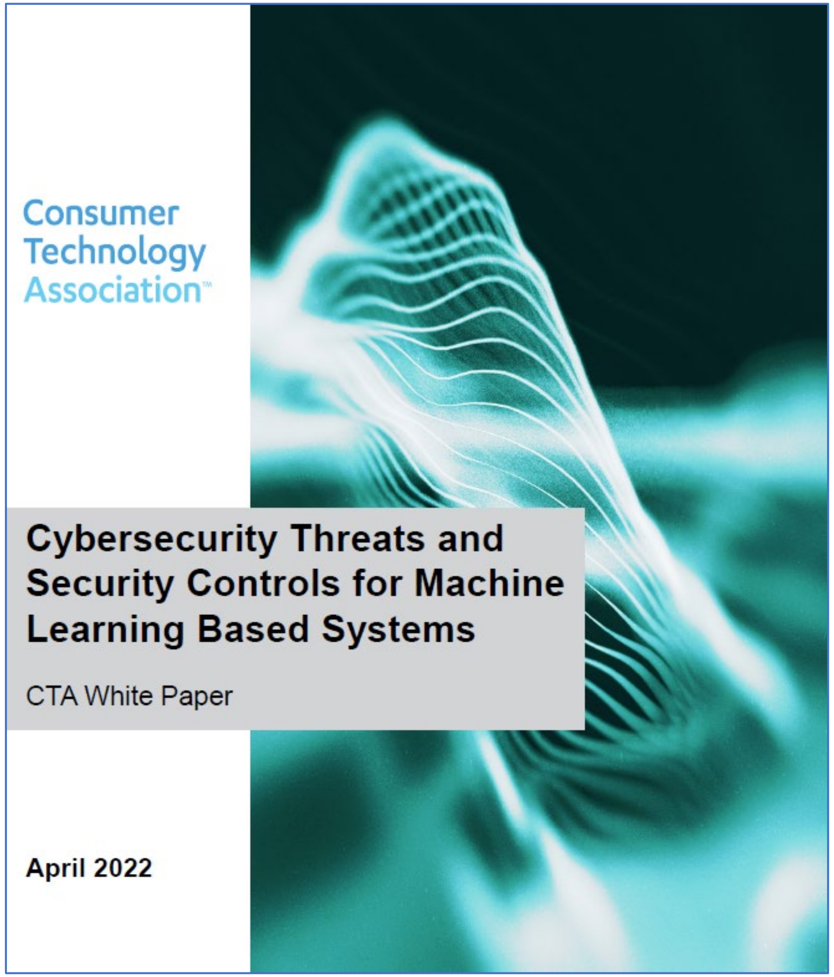
**Check the vulnerabilities of the components**
- Regular security audits, vulnerabilities scans, automatic patch management
- Impact on the availability of the system

**Monitor the performance of the model**
- Ensure the reliability of the model, be sure of the intelligence of the model by selecting quality data, always train and evaluate the model
- Improve the performance of the model

**Minimize data at each step of the processing**
- Study the necessity of collecting data such as age and body weight, proof the necessity of collecting such data
- Impact on security in case of forensic analysis

# Short whitepaper about this topic..

If you don’t have time to read 50 – 75 page reports on this topic, there is a 15 page whitepaper that provides an overview from CTA



Consumer Technology Association™

**Cybersecurity Threats and Security Controls for Machine Learning Based Systems**

CTA White Paper

**April 2022**

Free!
[https://shop.cta.tech/products/cybersecurity-threats-and-security-controls-for-machine-learning-based-systems-pdf](https://shop.cta.tech/products/cybersecurity-threats-and-security-controls-for-machine-learning-based-systems-pdf)

# CTA Paper continued..

- Wide variety of actors – creating data, collecting data, distributing data, cleaning, annotating, etc..
- Threats include:
- Poisoning
- Label modification
- Data or model disclosure
- Compromise of components (e.g. algorithms in an open-source library)

# CTA Paper continued..

Threats also extend to the post-deployment phase

- Evasion attacks – finding small perturbations that lead to significant actions (e.g. projecting images on the road causes autonomous vehicle to suddenly brake)
- Attacker puts a series of calculated inputs to the model and observe the outputs (reverse engineering) so they can do a better job in future attacks.
    - Could also use this technique to extract information about the data used to train the model.
- "Sponge Attacks" – a type of denial of service that takes advantage of the massive computational time that AI systems need by submitting malicious inputs that take extra time to compute. An example would be image recognition for self-driving cars.

# CTA Paper continued..

Possible controls include

- Intentionally including adversarial data – add adversarial examples throughout the lifecycle to make system more resilient to those attacks.
- Introduce random note into the dataset
- Use large datasets
- Continuous Poisoning Controls – not just during development, but after product launch.
- Federated Learning – try to have model training on multiple services that consist of local data samples (minimize sharing data with third parties)

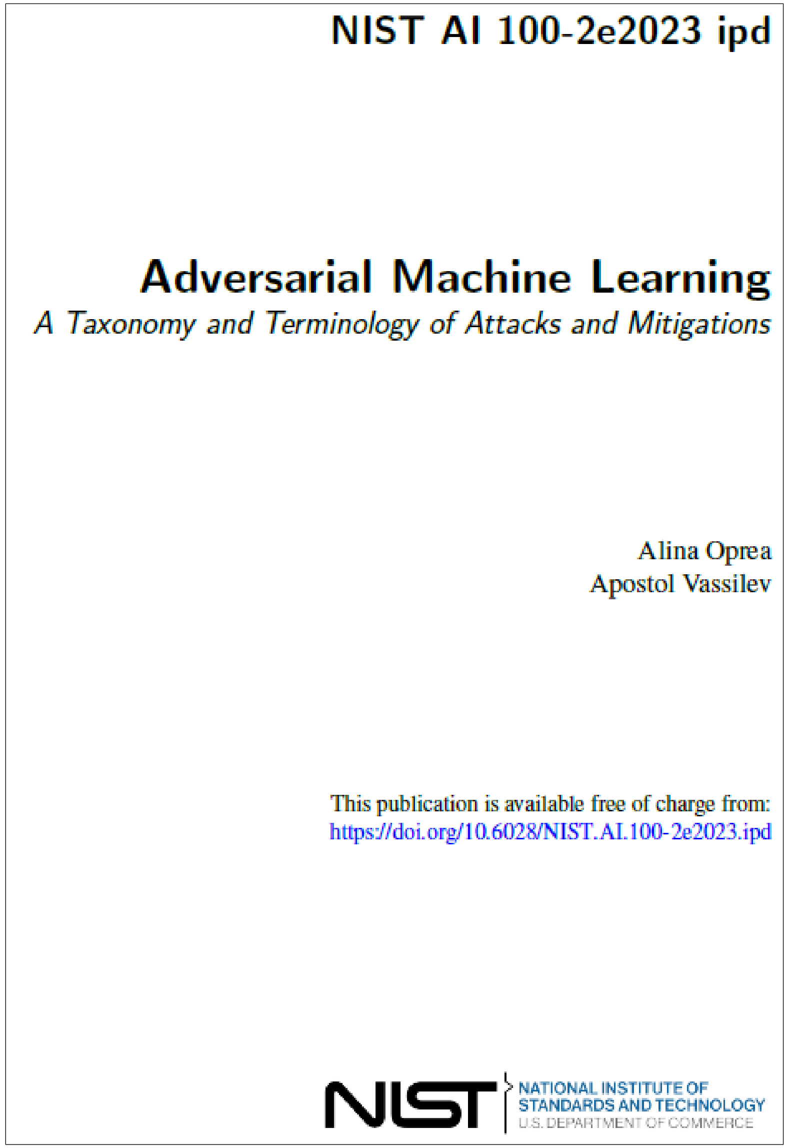
# Cybersecurity & Standards

- Standards organizations are starting to recognize that this is an issue and have started some projects to address it.
- ISO/IEC SC27 & SC42 have joined forces to work on an international standard for security of ML systems; however, this is a horizontal standard across all sectors – I am hoping that we can have an informative annex that addresses some of the unique considerations that we have in healthcare.
- CTA is currently working on “CTA-2114 Mitigating Cybersecurity Threats in ML-Based Systems”

# NIST AI (Draft) March 2023

NIST has published a 74-page draft list of attacks and mitigations. This includes:

- Attack Classification
- Evasion Attacks and Mitigations
- Poisoning Attacks and Mitigations
- Privacy Attacks
- Discussion and Remaining Challenges



NIST AI 100-2e2023 ipd

Adversarial Machine Learning

*A Taxonomy and Terminology of Attacks and Mitigations*

Alina Oprea
Apostol Vassilev

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.100-2e2023.ipd

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

# Other industries are working on Best Practices as Well

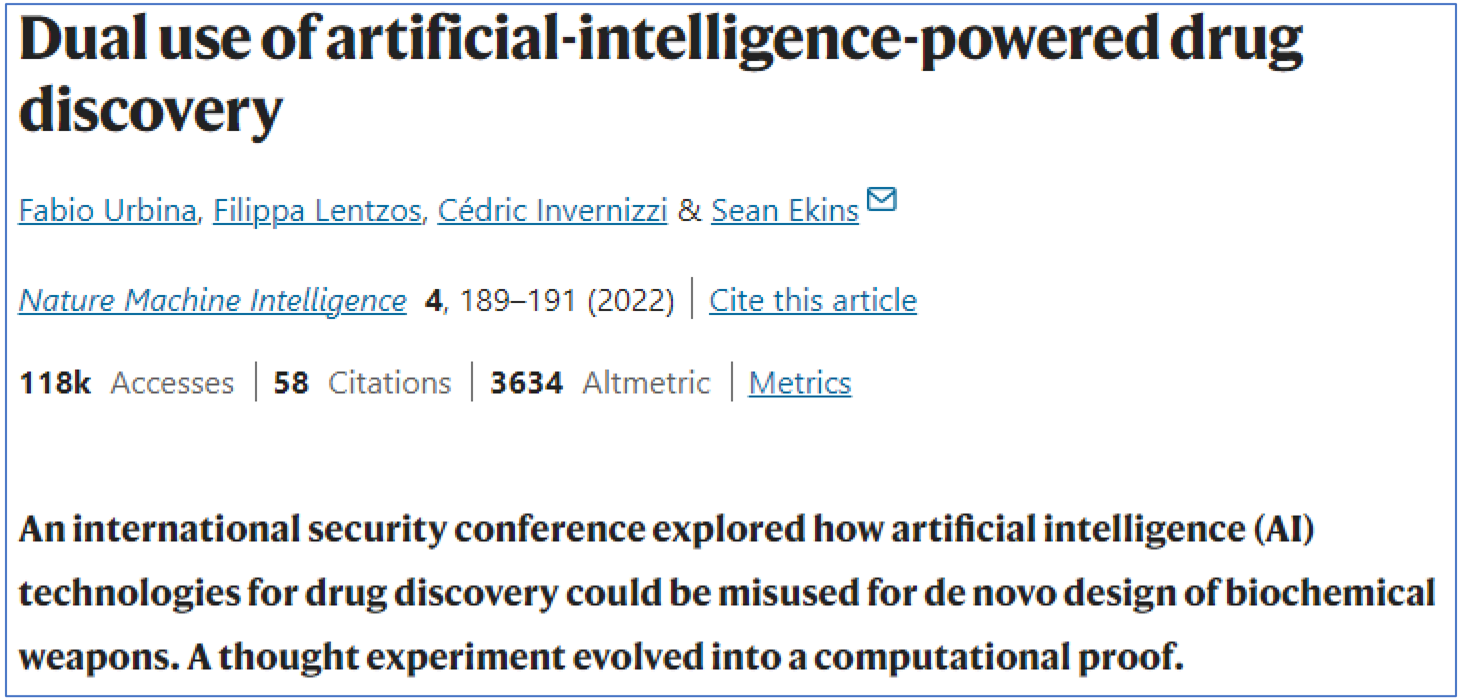
Cybersecurity is a concern for all industries & we can leverage the work of others. For example, while doing a literature search, I found that the US Department of Education setup a Privacy Technical Assistance Center to address privacy, confidentiality, and security practices. Although not specific to AI, there are some good tips regarding data destruction. Obviously, there are other standards and guidance regarding privacy & data destruction; my point is that we might want to look outside of our industry for ideas rather than starting from scratch..

http://ptac.ed.gov.

# Side topic: AI for Evil

AI improves our natural abilities in many ways, some of which can be used in negative ways.

This is an interesting article regarding how AI for drug discovery can be used to develop very effective bio-weapons.



## Dual use of artificial-intelligence-powered drug discovery

Fabio Urbina, Filippa Lentzos, Cédric Invernizzi & Sean Ekins

*Nature Machine Intelligence* **4**, 189–191 (2022) | Cite this article

**118k** Accesses | **58** Citations | **3634** Altmetric | Metrics

**An international security conference explored how artificial intelligence (AI) technologies for drug discovery could be misused for de novo design of biochemical weapons. A thought experiment evolved into a computational proof.**

https://www.nature.com/articles/s42256-022-00465-9

# Questions?