

# On Certification of Artificial Intelligence Systems

Dmitry Namiot, Eugene Ilyushin

Lomonosov Moscow State University  
dnamiot@gmail.com

GRID 2023

# Outline

- AI systems hereinafter - machine learning systems. Sometimes - deep learning systems. There is also AGI, but practically now it is reinforcement learning and LLM
- ML systems in the inference phase are programs
- Programs in critical applications are certified to prove operability.
- How to be in this case with ML applications?

# Content

- Introduction. Where does the proof-of-work problem for ML systems come from
- Legal regulations
- Audit of machine learning systems
- Certification of machine learning systems
- Technical basis for certification
- Discussion

# Introduction

## Three commandments of Secure/Safe ML

I. Thou shall not train on data you don't fully trust

(because of data poisoning)

II. Thou shall not let anyone use your model (or observe its outputs) unless you completely trust them

(because of model stealing and black box attacks)

III. Thou shall not fully trust the predictions of your model

(because of adversarial examples)

# Introduction

- NIST, according to the latest recommendations, distinguishes three basic types of attacks against machine learning systems:
- poisoning,
- evasion,
- attacks on intellectual property.
- The latter are a special survey of models in order to extract non-public information and do not affect the results of the work

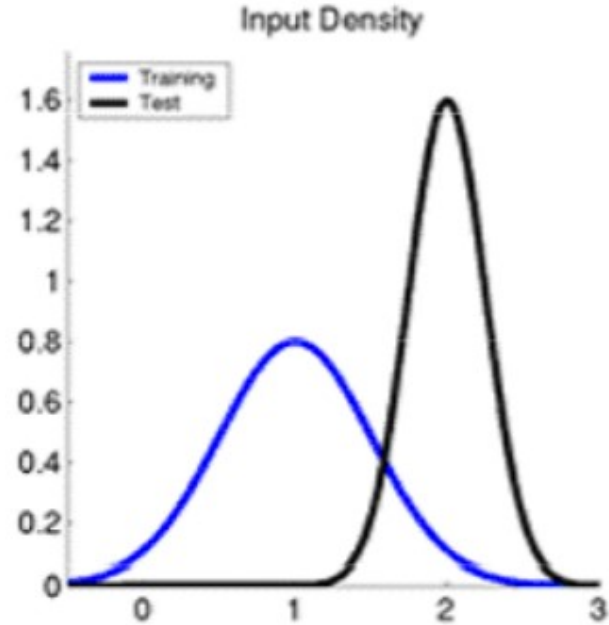
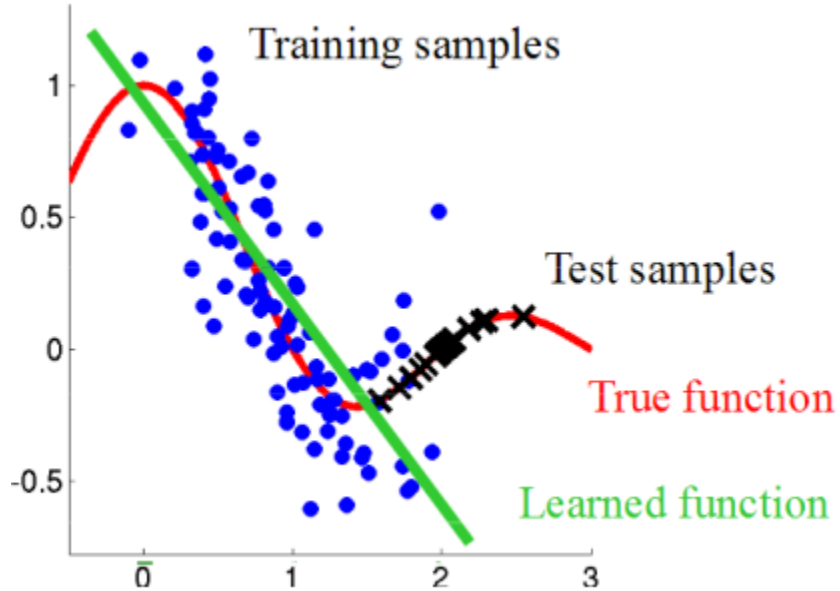
# Introduction

- The term poisoning is used to emphasize the long-term nature of the impact on models and includes data poisoning (special data modifications at the training stage) and model poisoning (direct modification of finished models).
- Such attacks require access to training data (or loading poisoned data) or loading modified (poisoned) models.
- In a first approximation, we can say that the requirements for protection against such attacks are similar to the usual requirements of cybersecurity (digital hygiene), with the prohibition of downloading anything from unknown sources (at least for critical applications this should definitely be excluded).

# Introduction

- What remains are evasion attacks, which consist in modifying (in the digital or physical domain) the input data.
- In the classical form, at the time of its appearance, these were the minimal modifications of the input data that caused the system to malfunction.
- For example, adding some noise to the correct data fools the classifier
- But: 1) the modified input data is the same data as all the rest. They cannot be detected by "antivirus" 2) misbehavior can happen without malicious actions
- Robustness – is a key !

# On the robustness



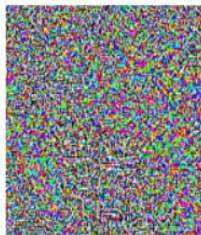


# On the evasion attacks



$x$   
“panda”  
57.7% confidence

+ .007 ×



$\text{sign}(\nabla_x J(\theta, x, y))$   
“nematode”  
8.2% confidence

=








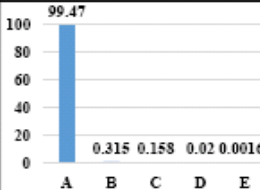
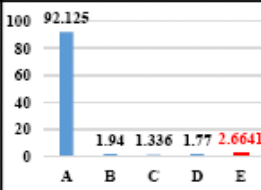
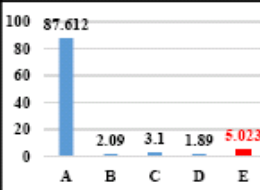
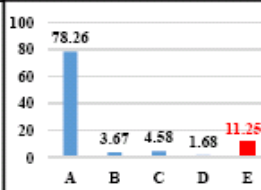
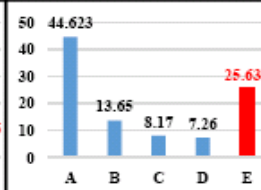
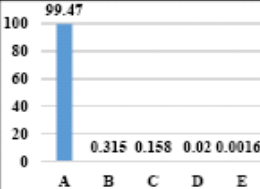
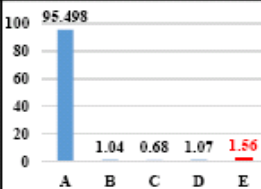
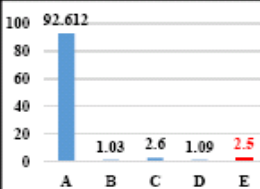
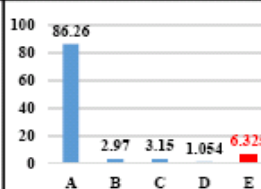
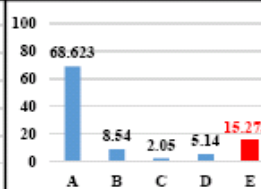
$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$   
“gibbon”  
99.3% confidence

- simple scheme
- exist for all discriminant models
- attacks always precede defenses
- explanations are different, but the nature is the same - we are only dealing with the part of the data during training.

# Summing up the problems

- We can get results with machine learning, but we can't guarantee them.
- Critical applications demand precisely guarantees
- As per DO-178C: "avionic systems should safely perform their intended function under all foreseeable operating and environmental conditions".
- Global evasion attacks can perturb any valid input example to mislead the model, whereas local evasion attacks can only perturb in-distribution data.
- Thus, the robustness should be hold for the whole input domain.

# Salt & Pepper attack

Noise	No Noise	1% Salt & Pepper	2% Salt & Pepper	3% Salt & Pepper	4% Salt & Pepper	Top5 Classes																																																												
Training Data Poisoning						A: Stop B: Yield C: Dangerous Curve to the Right D: Priority Road E: Speed Limit (60km/h)																																																												
Attack 1 Intrude the certain number of samples in training dataset	 <table border="1"> <tr><th>Class</th><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td></tr> <tr><th>Percentage</th><td>99.47</td><td>0.315</td><td>0.158</td><td>0.02</td><td>0.0016</td></tr> </table>	Class	A	B	C	D	E	Percentage	99.47	0.315	0.158	0.02	0.0016	 <table border="1"> <tr><th>Class</th><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td></tr> <tr><th>Percentage</th><td>92.125</td><td>1.94</td><td>1.336</td><td>1.77</td><td>2.6641</td></tr> </table>	Class	A	B	C	D	E	Percentage	92.125	1.94	1.336	1.77	2.6641	 <table border="1"> <tr><th>Class</th><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td></tr> <tr><th>Percentage</th><td>87.612</td><td>2.09</td><td>3.1</td><td>1.89</td><td>5.023</td></tr> </table>	Class	A	B	C	D	E	Percentage	87.612	2.09	3.1	1.89	5.023	 <table border="1"> <tr><th>Class</th><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td></tr> <tr><th>Percentage</th><td>78.26</td><td>3.67</td><td>4.58</td><td>1.68</td><td>11.266</td></tr> </table>	Class	A	B	C	D	E	Percentage	78.26	3.67	4.58	1.68	11.266	 <table border="1"> <tr><th>Class</th><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td></tr> <tr><th>Percentage</th><td>44.623</td><td>13.65</td><td>8.17</td><td>7.26</td><td>25.634</td></tr> </table>	Class	A	B	C	D	E	Percentage	44.623	13.65	8.17	7.26	25.634	Experimental Setup
Class	A	B	C	D	E																																																													
Percentage	99.47	0.315	0.158	0.02	0.0016																																																													
Class	A	B	C	D	E																																																													
Percentage	92.125	1.94	1.336	1.77	2.6641																																																													
Class	A	B	C	D	E																																																													
Percentage	87.612	2.09	3.1	1.89	5.023																																																													
Class	A	B	C	D	E																																																													
Percentage	78.26	3.67	4.58	1.68	11.266																																																													
Class	A	B	C	D	E																																																													
Percentage	44.623	13.65	8.17	7.26	25.634																																																													
Attack 2 Append the certain number of intruded samples in training dataset	 <table border="1"> <tr><th>Class</th><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td></tr> <tr><th>Percentage</th><td>99.47</td><td>0.315</td><td>0.158</td><td>0.02</td><td>0.0016</td></tr> </table>	Class	A	B	C	D	E	Percentage	99.47	0.315	0.158	0.02	0.0016	 <table border="1"> <tr><th>Class</th><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td></tr> <tr><th>Percentage</th><td>95.498</td><td>1.04</td><td>0.68</td><td>1.07</td><td>1.56</td></tr> </table>	Class	A	B	C	D	E	Percentage	95.498	1.04	0.68	1.07	1.56	 <table border="1"> <tr><th>Class</th><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td></tr> <tr><th>Percentage</th><td>92.612</td><td>1.03</td><td>2.6</td><td>1.09</td><td>2.5</td></tr> </table>	Class	A	B	C	D	E	Percentage	92.612	1.03	2.6	1.09	2.5	 <table border="1"> <tr><th>Class</th><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td></tr> <tr><th>Percentage</th><td>86.26</td><td>2.97</td><td>3.15</td><td>1.054</td><td>6.325</td></tr> </table>	Class	A	B	C	D	E	Percentage	86.26	2.97	3.15	1.054	6.325	 <table border="1"> <tr><th>Class</th><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td></tr> <tr><th>Percentage</th><td>68.623</td><td>8.54</td><td>2.05</td><td>5.14</td><td>15.273</td></tr> </table>	Class	A	B	C	D	E	Percentage	68.623	8.54	2.05	5.14	15.273	Dataset: German Traffic Sign Recognition Benchmarks Network: VGGNet Noise: Salt & Pepper # of Samples: 50000 # of Classes: 40
Class	A	B	C	D	E																																																													
Percentage	99.47	0.315	0.158	0.02	0.0016																																																													
Class	A	B	C	D	E																																																													
Percentage	95.498	1.04	0.68	1.07	1.56																																																													
Class	A	B	C	D	E																																																													
Percentage	92.612	1.03	2.6	1.09	2.5																																																													
Class	A	B	C	D	E																																																													
Percentage	86.26	2.97	3.15	1.054	6.325																																																													
Class	A	B	C	D	E																																																													
Percentage	68.623	8.54	2.05	5.14	15.273																																																													

# On regulations

- MIT Technology Review: “Suddenly, everyone wants to talk about how to regulate AI”.
- This applies to states, public associations (EU, G7), and even private companies (OpenAI, Google, Microsoft).
- Algorithmic Accountability Act (US), American Data Privacy Protection Act (US), ECAT - European Center for Algorithmic Transparency (EU), Measures for Generative Artificial Intelligence Services (China), The Artificial Intelligence Act (EU)
- Regulations describe the final state of products. There is no procedure for reaching such states

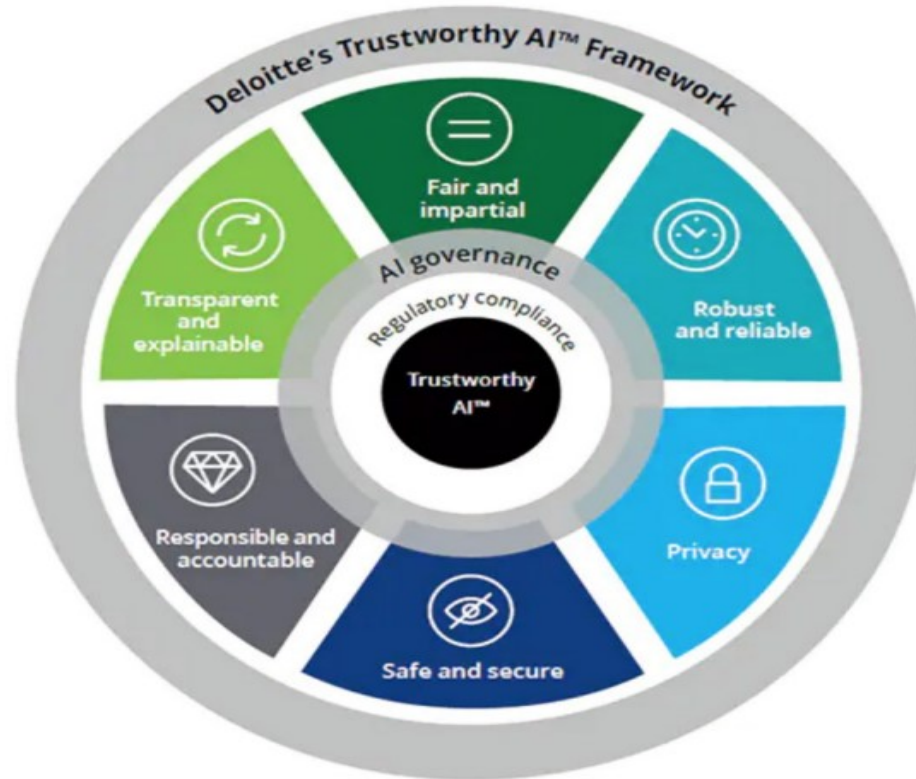
# On AI audit

- Classically: an audit is a process of inspection (verification), and certification is already a confirmation (guarantee) of data (work results).
- Auditing machine learning systems is a new and fairly rapidly developing area.
- Game Changers report 2022 (CB Insights) lists AI audit as the number one among 9 technologies that will change every industry
- Audit in one sentence: checklist. What developers/auditors should check/describe

# On AI audit areas

- Risk assessment before system deployment
- Hazard Opportunity Ratings
- Audit of third-party models
- Security testing (red team)
- Security restrictions
- Model Verification Techniques
- Security Incident Response Plan
- Pre-training risk assessment
- Monitoring systems and their use
- Model evaluations after deployment

# On AI audit frameworks



# On AI audit frameworks

Scoping	Mapping	Artifact Collection	Testing	Reflection	Post-Audit
Define Audit Scope	Stakeholder Buy-In	Audit Checklist	Review Documentation	Remediation Plan	Go / No-Go Decisions
Product Requirements Document (PRD)	Conduct Interviews	Model Cards	Adversarial Testing	Design History File (ADHF)	Design Mitigations
AI Principles	Stakeholder Map	Datasheets	Ethical Risk Analysis Chart		Track Implementation
Use Case Ethics Review	Interview Transcripts			Summary Report	
Social Impact Assessment	Failure modes and effects analysis (FMEA)				



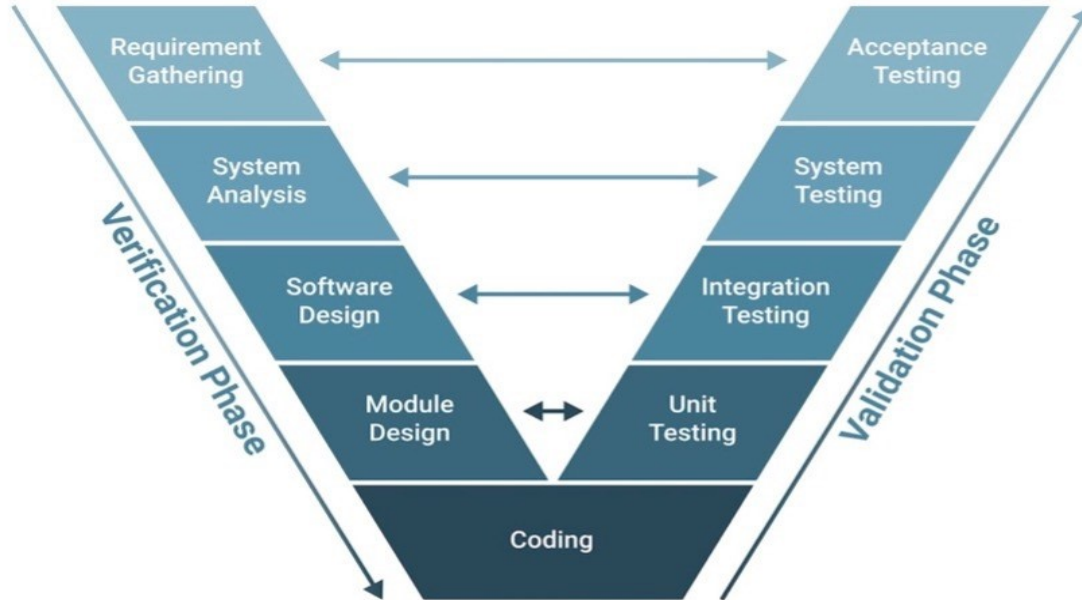
# On AI audit frameworks

- NIST AI RISC Management framework
- ISO/IEC 23894
- Fraunhofer
- Gartner AI TRiSM (Artificial Intelligence Trust, Risk, and Security Management)
- In terms of implementations, this includes AI trusted platforms. E.g.: Datarobot, IBM Trustworthy

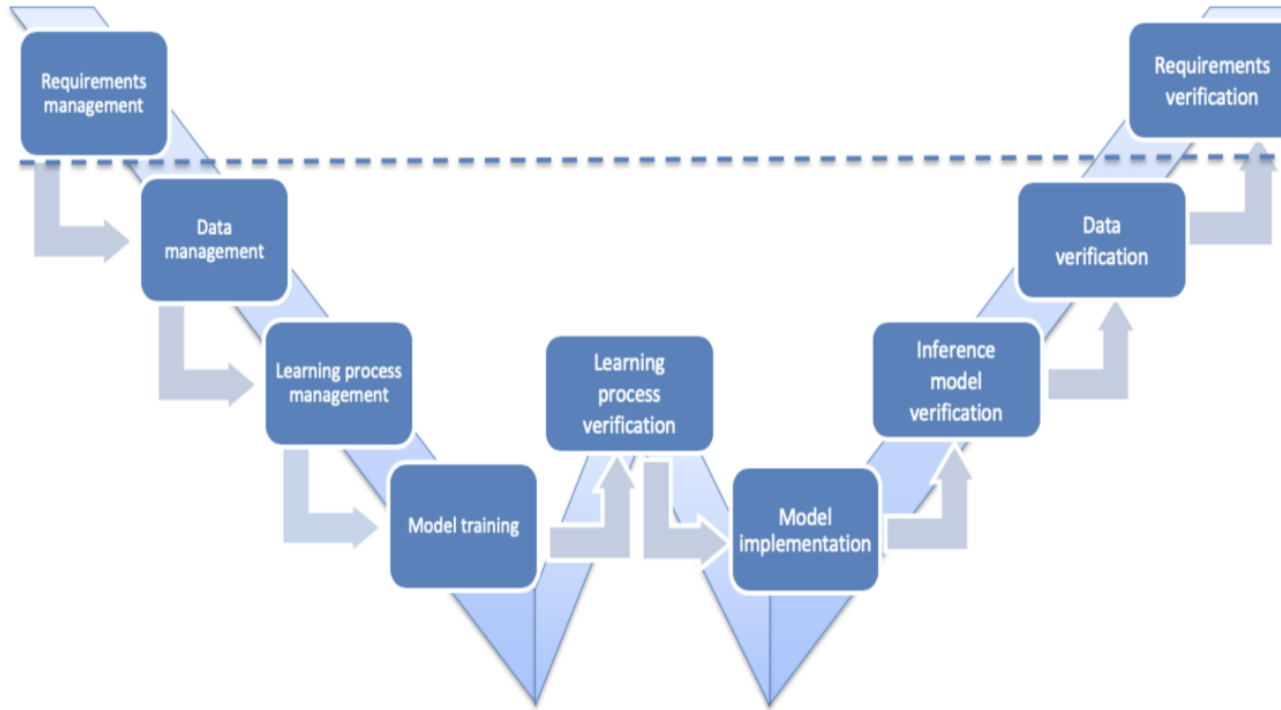
# On certification

- Software Assurance (SwA) is a critical software development process that ensures that software products are reliable, secure, and safe.
- It includes many activities: requirements analysis, design analysis, code review, testing, and formal review.
- One of the most important components of software security is secure coding practices that comply with industry standards and best practices.
- V-model:
  - Verification – are we building the product correctly?
  - Validation - is the correct product built

# On certification



# On certification



# Main inconsistencies

- deterministic approach in certification of software systems against non-deterministic ML models
- code coverage (it is based on V-model - why is this line in the code?)
- data coverage. A standard approach in ML – point-wise robustness. Certification for ML models is a study of robustness in some limited range of modifications of correct data. E.g. 35.42% certified accuracy on MNIST under perturbation  $\epsilon = 8/255$ .

# EASA roadmap

## Level 1 AI: assistance to human

- Level 1A: Human augmentation
- Level 1B: Human cognitive assistance in decision-making and action selection

## Level 2 AI: human-AI teaming

- Level 2A: Human and AI-based system cooperation
- Level 2B: Human and AI-based system collaboration

## Level 3 AI: advanced automation

- Level 3A: The AI-based system performs decisions and actions that are overridable by the human.
- Level 3B: The AI-based system performs non-overridable decisions and actions (e.g. to support safety upon loss of human oversight).

# Technical basis for certification

- The robustness verification: to evaluate robustness by providing a theoretically certified lower bound of robustness *under certain perturbation constraints*;
- The robust training: to train networks to improve such lower bound
- Complete verification: when the verification approach outputs “not verified” for a given  $x_0$ , if it is guaranteed that an adversarial example  $x$  around  $x_0$  exists; and otherwise incomplete verification.
- Deterministic verification: when the given input is non-robust against the attack, is guaranteed to output “not verified”; and the *probabilistic verification*: output “not verified” with a certain probability (e.g., 99.9%) where the randomness is independent of the input.

# Discussion

- Legal regulation of AI is not relevant to the proof of workability of AI systems
- Auditing AI systems is a practical and feasible step that should be applied to all industrial systems. EASA concept paper: “First usable guidance for level 1 machine learning applications” can be recommended as a basis for audit (corporate, industry or national standards).
- Certification in the classical form is not possible (yet?) . The only approach corresponding to the classical model is formal verification for machine learning systems.