

## Invited Talk

### Assessing machine learning systems from a security and trust standpoint



Robson de Oliveira Albuquerque  
University of Brasília

Faculty of Computer Science and Engineering, UCM  
Madrid (Spain), July 2024

## Disclaimer

**The content presented herein represents my thoughts about the subject of this lecture.**

**It does not necessarily represent the vision of my employers nor the University of which I am a researcher;**

**Images used in this presentation are the property and credit of the respective creators and sources;**

## AGENDA

Machine Learning in the past;

Machine Learning nowadays;

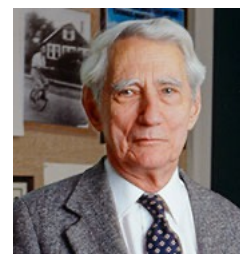
Security and trust;

Attacks on AI systems;

Discussions;

## Machine Learning in the past

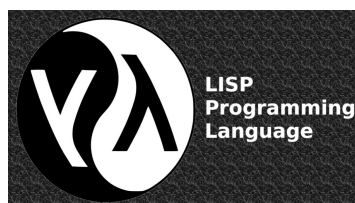
- Before 1950 – Statical Methods research
- 1950 – The real start...
  - The Dartmouth Summer Research Project on Artificial Intelligence in 1956 summer workshop - artificial intelligence as a field.
  - John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon.



## Machine Learning in the past

### • 1950 – 1960

- LISP language (great influence in other languages)



## Machine Learning in the past

### • 1960 – 1970

- Bayesian methods are introduced for **probabilistic inference** in machine learning
  - using data analysis to infer properties of an underlying distribution of probability
  - basically, is to update the probability for a hypothesis as more evidence or information becomes available

## Machine Learning in the past

- **1970 – 1980 – AI winter**
  - **Research interest drops significantly;**
  - **1971–75 - DARPA's frustration with the Speech Understanding Research program at Carnegie Mellon**
  - **1973 - Lighthill, James. "Artificial Intelligence: A General Survey". Artificial Intelligence: A paper symposium. UK: Science Research Council.**
  - **1973–74 - DARPA's cutbacks to academic AI research**

## Machine Learning in the past

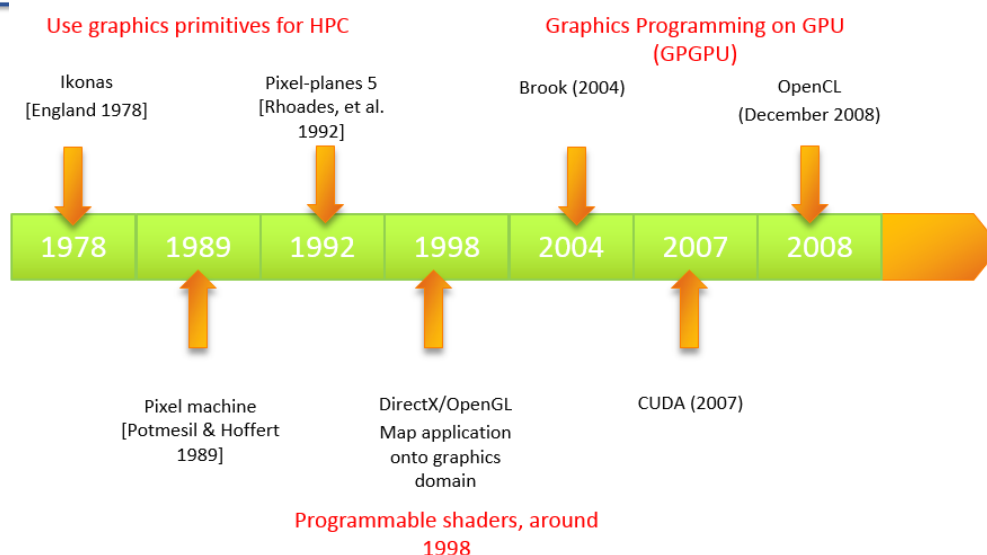
- **1980 – 1990**
  - **Rediscovery of **backpropagation** causes a resurgence (it was created in the 60s);**
    - **method used to train neural network models;**
  - **1987 - LISP machine market gets collapsed...**

## Machine Learning in the past

### • 1990 – 2000

- shifts from a knowledge-driven approach to a data-driven approach;
- programs for computers to analyze large amounts of data and draw conclusions – or "learn" – from the results;
- Support-vector machines (SVMs) and recurrent neural networks (RNNs) become popular.

## Machine Learning in the past (GPUs)



## Machine Learning in the past

- **2000 – 2010**

- **Methods like Support-Vector Clustering starts...**
- **unsupervised machine learning methods become widespread...**

- **2010 – 2020**

- **Deep Learning (spurs huge advances in vision and text processing)**
- **Generative AI**

## AGENDA

Machine Learning in the past;











Machine Learning nowadays;

Security and trust;

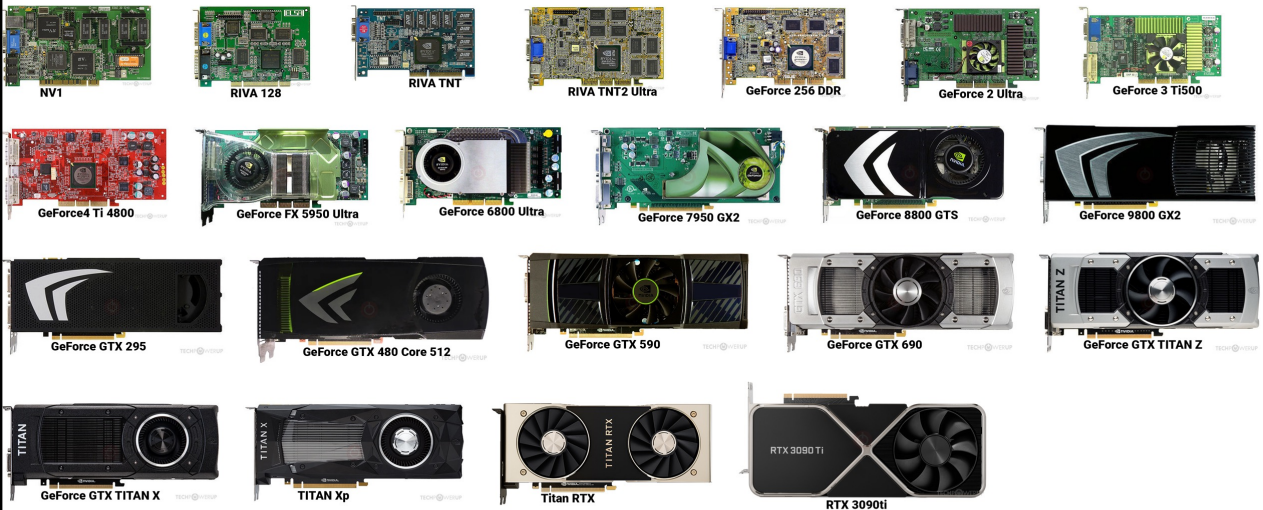
Attacks on AI systems;

Discussions;

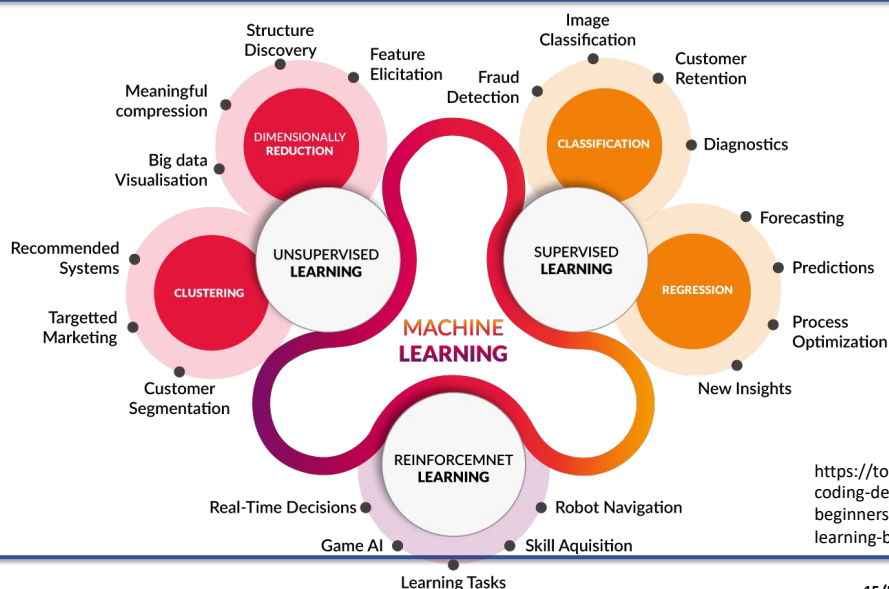
# Machine Learning nowadays

 2008	GTX 280 MSRP: \$649 (\$899 Adjusted for inflation)	 2016	GTX 1080 MSRP: \$599 (\$745 Adjusted for inflation)
 2010	GTX 480 MSRP: \$499 (\$680 Adjusted for inflation)	 2018	RTX 2080 MSRP: \$699 (\$830 Adjusted for inflation)
 2011	GTX 580 MSRP: \$499 (\$660 Adjusted for inflation)	 2020	RTX 3080 MSRP: \$699 (\$800 Adjusted for inflation)
 2012	GTX 680 MSRP: \$499 (\$649 Adjusted for inflation)	 2022	RTX 4080 MSRP: \$1199
 2013	GTX 780 MSRP: \$649 (\$830 Adjusted for inflation)		
 2014	GTX 980 MSRP: \$549 (\$690 Adjusted for inflation)		

# Machine Learning nowadays



## Machine Learning nowadays



## Machine Learning nowadays

- Ok...
  - Everything is cool...
  - Lot's of papers/advances...
  - Lot's of people says they are "masters of the art" ...
  - Lot's of money in Start-ups...
  - A lot of lot's...



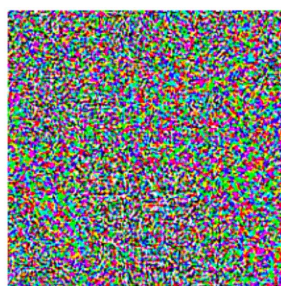
## Machine Learning nowadays

- We saw things like this few years ago...



“panda”  
57.7% confidence

+ .007 ×



“nematode”  
8.2% confidence

=



“gibbon”  
99.3 % confidence

## Machine Learning nowadays

- We saw things like this few years ago...

**AI within  
cybersecurity**

	Malware	Score Before	Score After
CoinMiner		-826	884
Dridex		-999	996
Emotet		-923	625
Gh0stRAT		-975	998
Kovter		-999	856
Nanobot		-971	999
Pushdo		-999	999
Qakbot		-998	991
Trickbot		-973	774
Zeus		-997	997

# Machine Learning nowadays

• We saw things like this few years ago...



# Machine Learning nowadays

• We saw things like this few years ago...

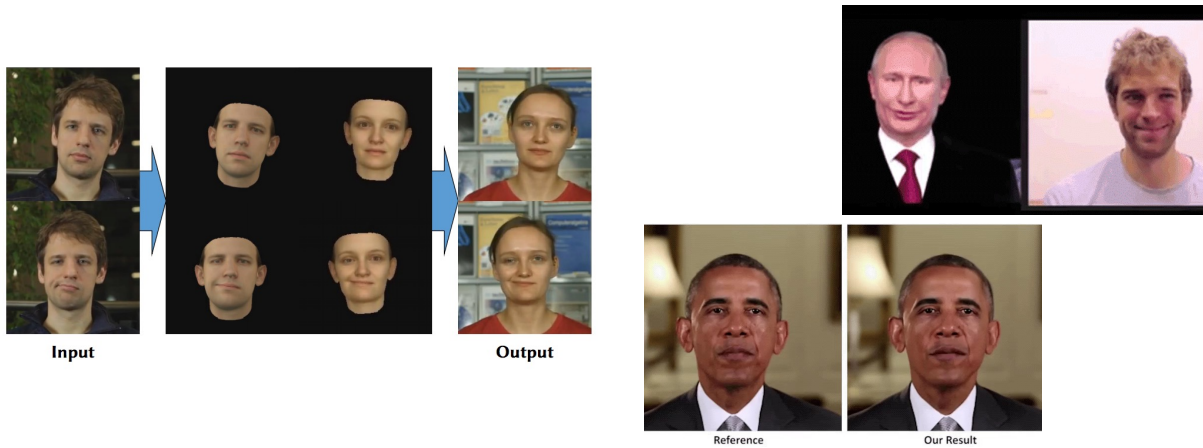
## Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men

Facebook called it "an unacceptable error." The company has struggled with other issues related to race.



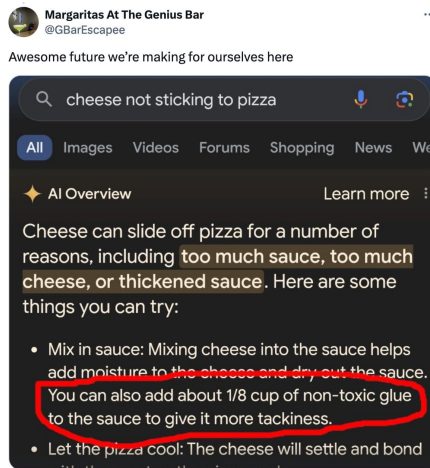
# Machine Learning nowadays

• We saw things like this few years ago...



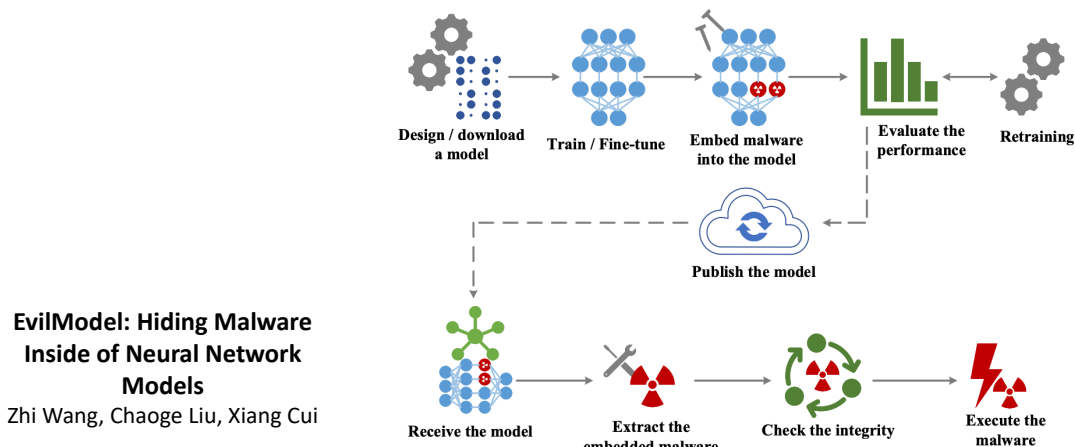
# Machine Learning nowadays

• We are seeing things like this these days...



## Machine Learning nowadays

• We are seeing things like this these days...



## Machine Learning nowadays

• We are seeing things like this these days...

### Data Scientists Targeted by Malicious Hugging Face ML Models with Silent Backdoor



By David Cohen, JFrog Senior Security Researcher | February 27, 2024

🕒 13 min read

SHARE:

## Machine Learning nowadays

• We are seeing things like this these days...

CRIME

### Deepfake scams have robbed companies of millions. Experts warn it could get worse

PUBLISHED MON, MAY 27 2024-10:20 PM EDT | UPDATED TUE, MAY 28 2024-5:31 AM EDT



Dylan Butts  
@IN/DYLAN-B-7A451A107

SHARE [f](#) [X](#) [in](#) [✉](#)

#### KEY POINTS

- Earlier this year, a Hong Kong finance worker was duped into transferring \$25 million to a fraudster that had deepfaked his chief financial officer and ordered the transfer via video call.



<https://www.cnn.com/2024/05/28/deepfake-scams-have-looted-millions-experts-warn-it-could-get-worse.html>

25/50

25

## Machine Learning nowadays

• We are seeing things like this these days...

### Hackers Can Use AI Hallucinations to Spread Malware

A Fake Software Library Made Up by a ChatBot Was Downloaded More Than 35,000 Times

Rashmi Ramesh ([@rashmiramesh\\_](#)) · April 5, 2024



<https://www.bankinfosecurity.com/hackers-use-ai-hallucinations-to-spread-malware-a-24793>

26/50

26

# Machine Learning nowadays

## • We are seeing things like this these days...

Sam Altman, chief executive of ChatGPT maker OpenAI, has warned people not to expect artificial intelligence to do all the work for them, saying the technology is “far from perfect” and does not offer a shortcut to building great businesses. Mr Altman made a surprise appearance at Microsoft annual Build developer conference in Seattle early Wednesday morning (Australian time), and offered a glimpse to where AI is headed and gave some advice to entrepreneurs. He was speaking after Microsoft chief executive Satya Nadella announced a suite of updates to the company’s Copilot AI platform, including a Teams Copilot that aims to lift productivity across departments or entire companies rather than individual users.

### OpenAI’s Altman warns on ‘losing sight in AI gold rush’



OpenAI CEO Sam Altman speaking at the Microsoft Build conference this week. Picture: Jason Redmond/AFP

By JARED LYNCH  
7:14AM MAY 22, 2024



<https://archive.is/47BCN>

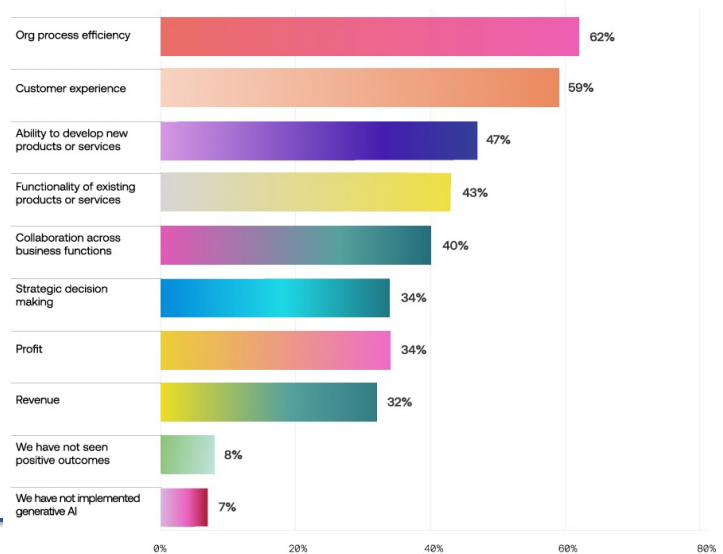
27

# Machine Learning nowadays

## • We are seeing things like this these days...

### • Scale's "2024 AI Readiness Report"

#### What positive outcomes have you seen from generative AI adoption?



<https://scale.com/ai-readiness-report#section-download>



28

## Machine Learning nowadays

### • Concerns:

- the quality and reliability of the AI-generated code;
- It will produce incorrect or inefficient code;
- it will suggest code snippets that may contain security vulnerabilities;
- Copyright of replicated code of private repositories used to train the model;

## AGENDA

Machine Learning in the past;

Machine Learning nowadays;

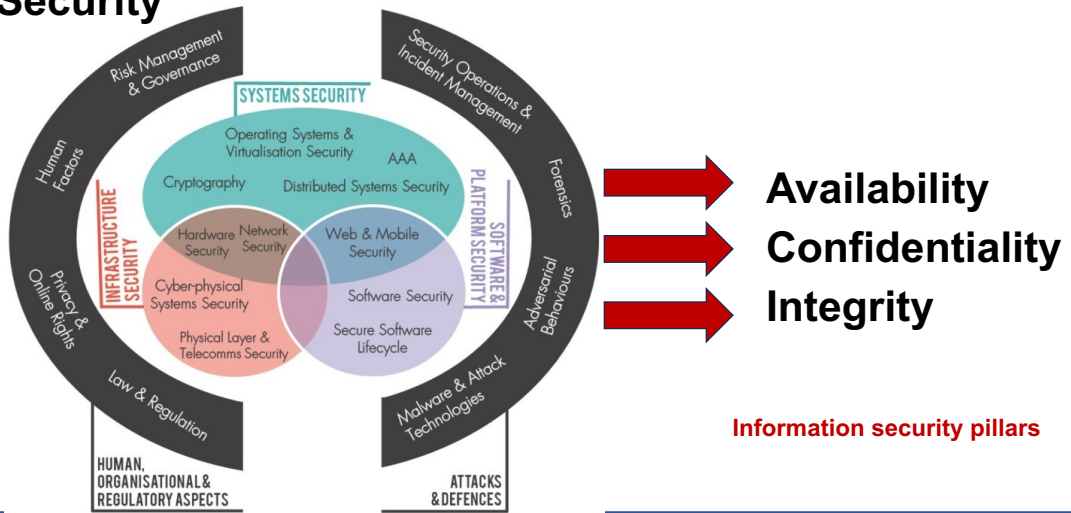
**Security and trust;**

Attacks on AI systems;

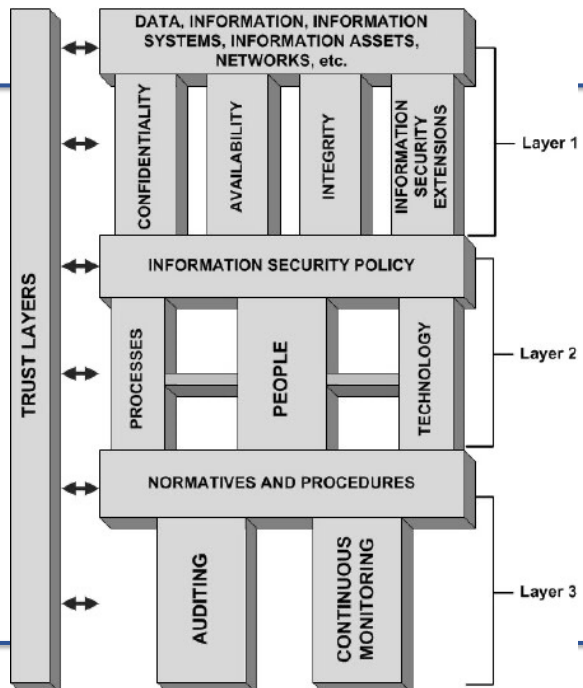
Discussions;

# Security and trust

## Cyber Security

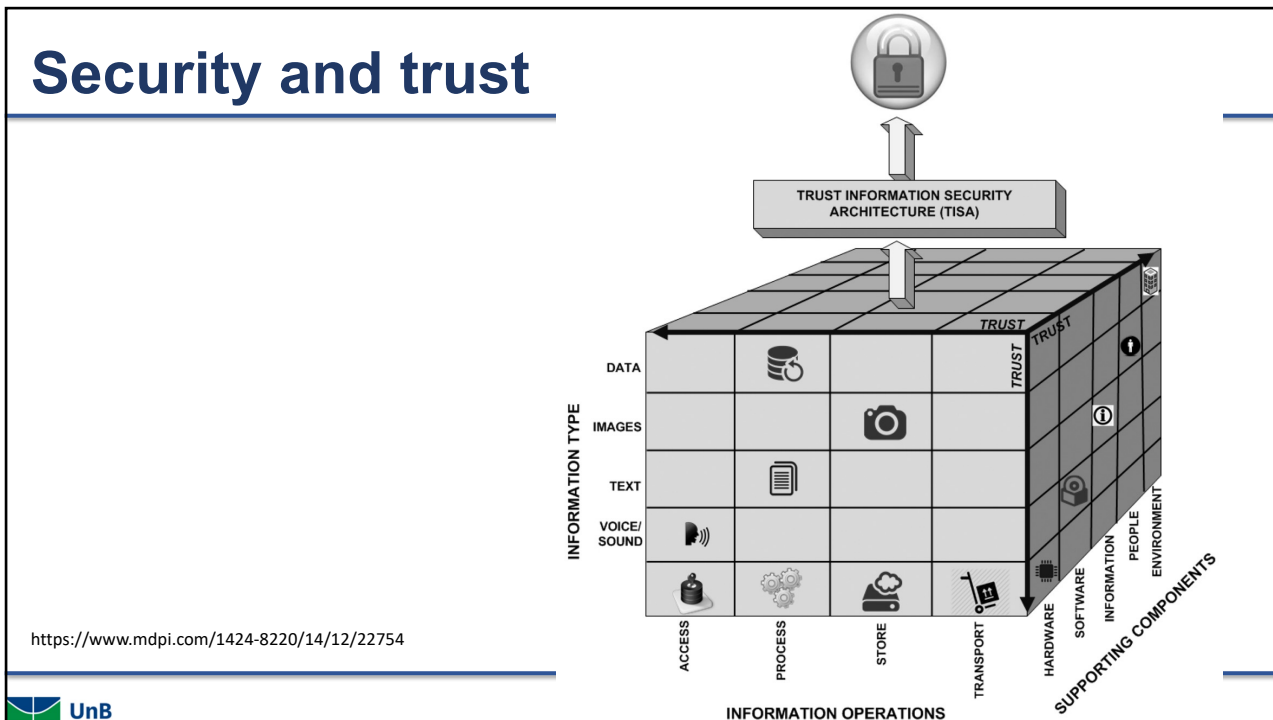


# Security and trust



<https://www.mdpi.com/1424-8220/14/12/22754>





33

# Security and trust

Google Scholar "trust in llms" About 99 results 0.05 sec

Articles

Any time  
 Since 2024  
 Since 2023  
**Since 2020**  
 Custom range...

Sort by relevance  
 Sort by date

Any type  
 Review articles

include patents  
 include citations  
 Create alert

**Enhancing Trust in LLMs: Algorithms for Comparing and Interpreting LLMs** [PDF] arxiv.org  
 NB Brown - arXiv preprint arXiv:2406.01943, 2024 - arxiv.org  
 This paper surveys evaluation techniques to enhance the trustworthiness and understanding of Large Language Models (LLMs). As reliance on LLMs grows, ensuring their reliability, ...  
 ☆ Save 📄 Cite Related articles 🔗


**Beyond algorithmic trust: interpersonal aspects on consent delegation to LLMs**  
 Z Sassi, M Hahn, S Eickmann... - Journal of Medical ..., 2024 - jme.bmj.com  
 ... Therefore, when discussing questions about **trust in LLMs** tasked with obtaining informed ... interpersonal aspects of trust in their reflections on **trust in LLMs**. With this in mind, we wanted ...  
 ☆ Save 📄 Cite Related articles All 5 versions

**Large Language Models and User Trust: Focus on Healthcare** [PDF] arxiv.org  
 A Choudhury, Z Chaudhry - arXiv preprint arXiv:2403.14691, 2024 - arxiv.org  
 ... This paper explores the evolving relationship between clinician **trust in LLMs**, the transformation of data sources from predominantly human-generated to AI-generated content, and the ...  
 ☆ Save 📄 Cite Cited by 2 Related articles All 2 versions 🔗

UnB 34/50

34

# Security and trust

Google Scholar "trust" "Large Language Model" "devsecops" 

Articles About 25 results (0.08 sec)



Any time  
 Since 2024  
 Since 2023  
Since 2020  
 Custom range...



Sort by **relevance**  
 Sort by date

Any type  
 Review articles


include patents  
 include citations

Create alert

**AI for DevSecOps: A Landscape and Future Opportunities** [PDF] arxiv.org  
 M Fu, J Pasuksmit, C Tantithamthavorn - arXiv preprint arXiv:2404.04839, 2024 - arxiv.org  
 ... DevSecOps paradigm simultaneously. This paper seeks to contribute to the critical intersection of AI and DevSecOps ... identifying avenues for enhancing security, **trust**, and efficiency in ...  
 ☆ Save  Cite Cited by 2 Related articles All 3 versions 

**The Convergence of AI/ML and DevSecOps: Revolutionizing Software Development** [PDF] jklst.org  
 N Pakalapati, S Venkatasubbu, SMK Sistla - Journal of Knowledge ..., 2023 - jklst.org  
 ... (ML) with DevSecOps represents a groundbreaking ... /ML technologies into the DevSecOps framework, revolutionizing ... for leveraging AI/ML in DevSecOps. Topics addressed include ...  
 ☆ Save  Cite Related articles 

**From LLMops to DevSecOps for GenAI**  
 K Huang, V Manral, W Wang - Generative AI Security: Theories and ..., 2024 - Springer  
 ... chooses an appropriate pre-trained **large language model** based on the specific needs of ... can lead to dissatisfaction or loss of **trust**. Ensuring availability requires careful planning, ...

 UnB 35/50

35

# AGENDA

---

Machine Learning in the past;

---

Machine Learning nowadays;

---

Security and trust;


---

Attacks on AI systems;

---

Discussions;

---

 UnB 36/50

36

## Attacks on AI systems

- **3 classifications are ordinary:**
  - **Adversarial inputs – Forged input data for misclassification;**
  - **Data poisoning attacks – Pollute classifier data in favor of the attacker;**
  - **Model stealing techniques – used to steal the model or retrieve/rebuild training data;**

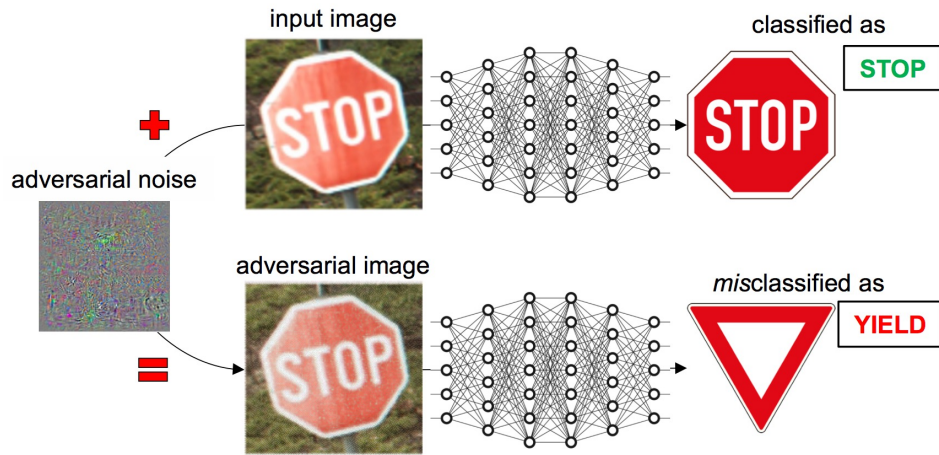
## Attacks on AI systems (adv. inputs)

autonomous driving systems



# Attacks on AI systems (adv. inputs)

autonomous driving systems



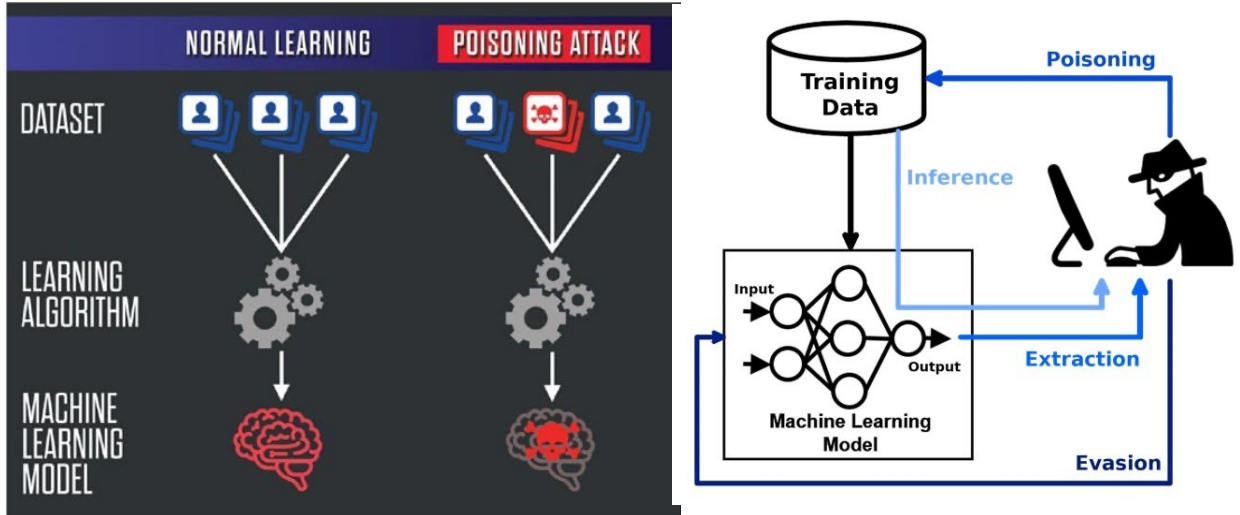
# Attacks on AI systems

**Autopilot sends Tesla onto train tracks.  
California cops say feature mistook  
railroad for street**

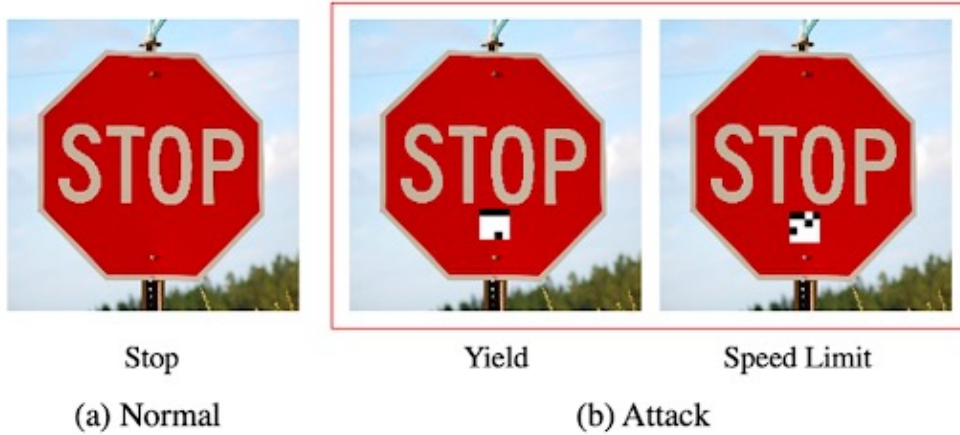
BY VINCENT MEDINA  
JUNE 28, 2024 1:31 PM



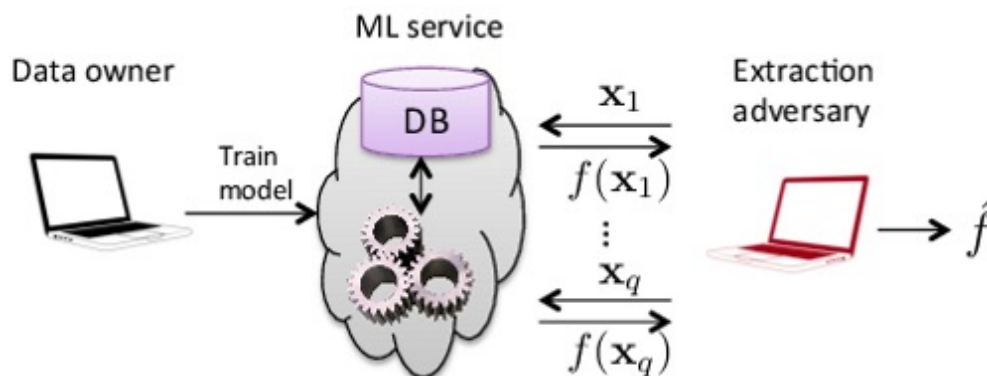
# Attacks on AI systems (data poison.)



# Attacks on AI systems (data poison.)



## Attacks on AI systems (model steal.)



## Attacks on AI systems (model steal.)

- Train models is a hard task these days...
  - Adversaries can break business models...
  - Leads to privacy issues...

## AGENDA

Machine Learning in the past;

Machine Learning nowadays;

Security and trust;

Attacks on AI systems;

**Discussions;**

## Discussions

- **AI with cybersecurity:**
  - **Malware or not malware?**
  - **Attack on the go or normal operations?**
- **AI within medicine:**
  - **Cancer or not cancer?**
- **AI with smart vehicles:**
  - **Hack or not Hack?**
  - **IA driver dilemma?**

## Discussions

- **In near future...**

- **AI is supposed to take decisions on our behalf...**
- **AI will need more and more data...what about privacy in LLMs?**
- **AI is supposed to classify speech...(good, bad, or what??) – will we trust LLM systems to tell us what if a text is fake or not?**

## Discussions

- **There is a huge avenue for research as well...**

- **So, do you trust your AI system?**
- **So, the model you created/downloaded is secure?**



# READY FOR...

**Q&A**



# Invited Talk

**Assessing machine learning systems from a security and trust standpoint**



**Robson de Oliveira Albuquerque**  
University of Brasília

**Faculty of Computer Science and Engineering, UCM**  
Madrid (Spain), July 2024