AI Challenges from Perspective of Cyber Security

Challenges overview, real life examples and future trends

Presentation made by "AIMS - Autonomne Inteligentne Mašine i Sistemi" and "Advanced Security Technologies"

ADVANCED SECURITY TECHNOLOGIES

MAJOR TOPICS TODAY

- Regulations changes because of AI solutions
- AI powered Cyber Attacks
- Attacks against AI solutions

ADVANCED SECURITY TECHNOLOGIES

MAJOR TOPICS TODAY

- **Regulations changes because of AI solutions**
- AI powered Cyber Attacks
- Attacks against AI solutions

AI Regulation Changes

On 14 June 2023, the European Parliament adopted its position on the AI Act. Parliament's priority is to make sure that AI systems used in the EU are **safe**, **transparent**, **traceable**, non-discriminatory and environmentally friendly.

Some impacts are:

- Transparent, traceable AI solution requirements:
  - Black Box AI solutions will be probably **banned for some services** like health, safety, fundamental rights, the environment, democracy and rule of law) -> some **Deep Learning Models or Complex Support Vector Machines (SVMs) will not be allowed for usage**
- **AI based service**/product providers must **register their AI models in the EU database before their release** on the EU market.

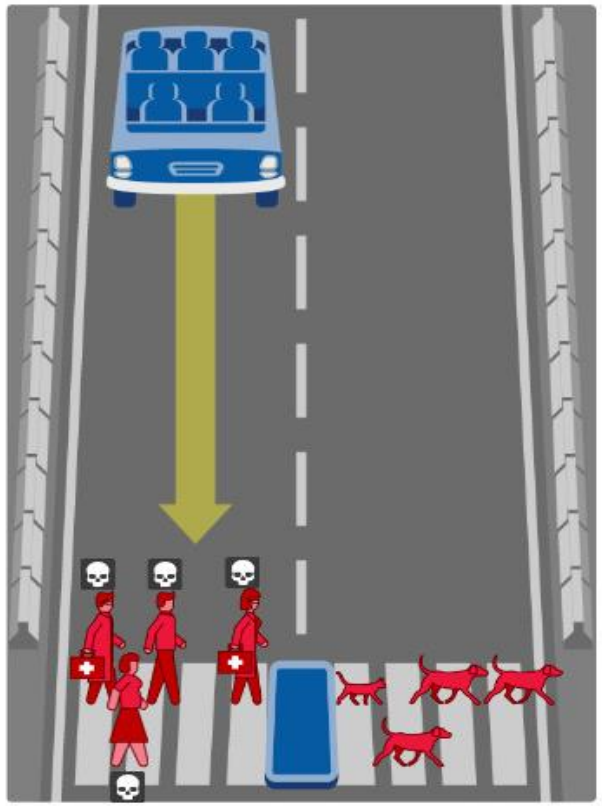
- EU Parliament proposed guidelines for use of AI, especially in areas such as military, justice and health that must not relieve humans of their responsibility

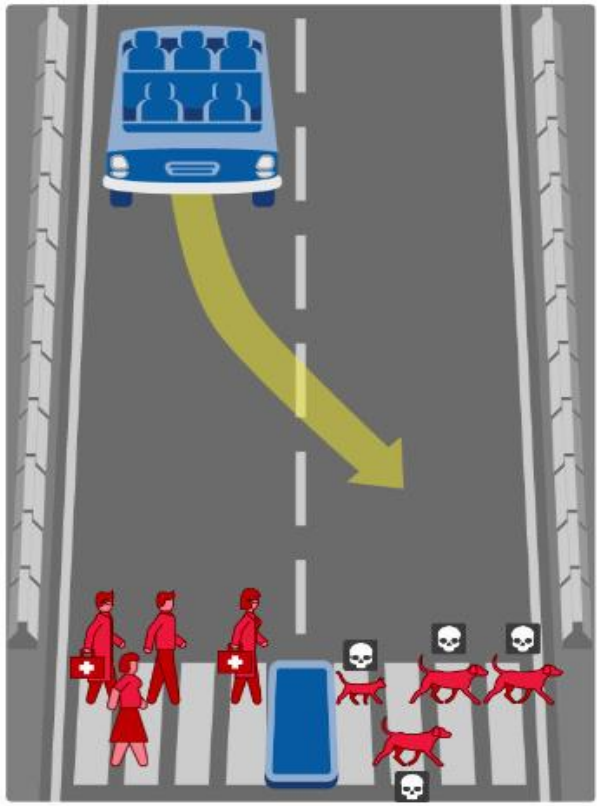
- Waiting for AI Law, expected in Dec 2023

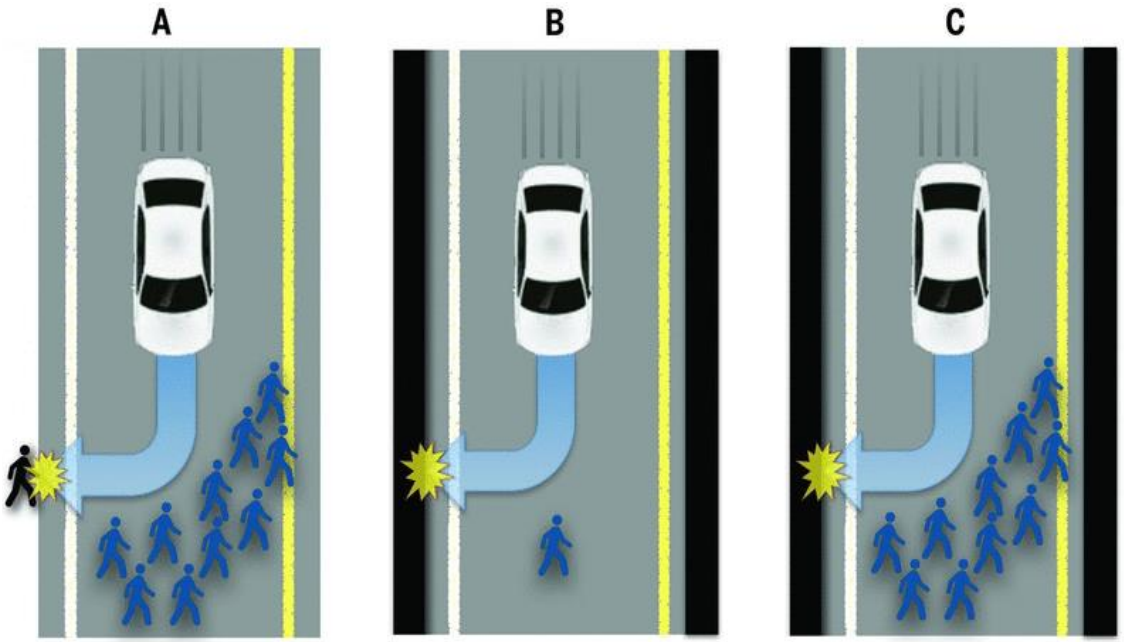
AI Regulation Changes

Who is responsible if the AI driven car hits a human on the road?

a) Human driver b) AI solution c) Software solution







Solution might be that everyone is assumed guilty unless proven opposite.

ADVANCED SECURITY TECHNOLOGIES

MAJOR TOPICS TODAY

- Regulations changes because of AI solutions
- **AI powered Cyber Attacks**
- Attacks against AI solutions

AI Powered Cyber Attack tools

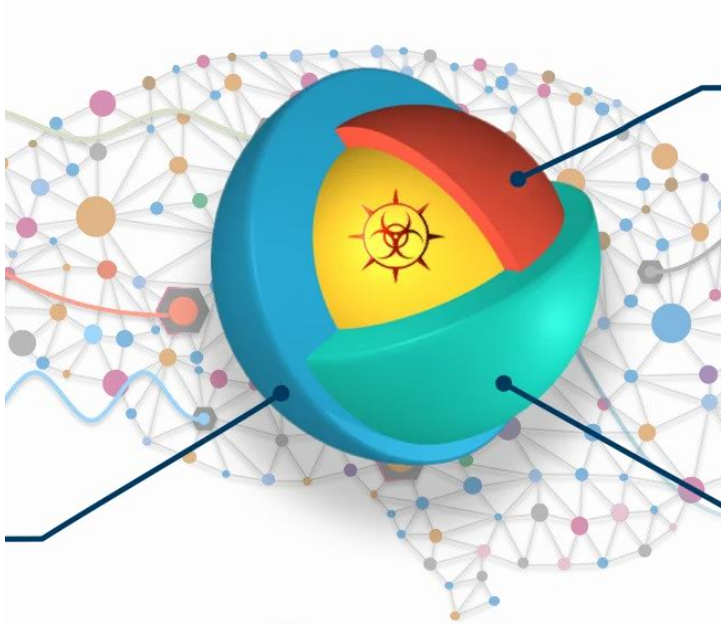
As soon as new AI solutions arrived,
AI based attack tools have been created too.

And we are at the very beginning

AI Powered Cyber Attack tools

Black Mamba

- large langage model (LLM)
- AI-generated polymorphic malware
- used for keylogging
- passed security solutions



DeepLocker

- AI makes the "trigger conditions" by using a deep neural network (DNN) model (visual, audio, geolocation, system-level features)
- Black box DNN hides the logic, avoiding rule based detection

WormGPT

- primarily writing effective phishing emails.
- can be used to write code automatically – including malware and cybersecurity exploits.
- Unrestricted AI capabilities
- (100-500USD)

ADVANCED SECURITY TECHNOLOGIES

MAJOR TOPICS TODAY

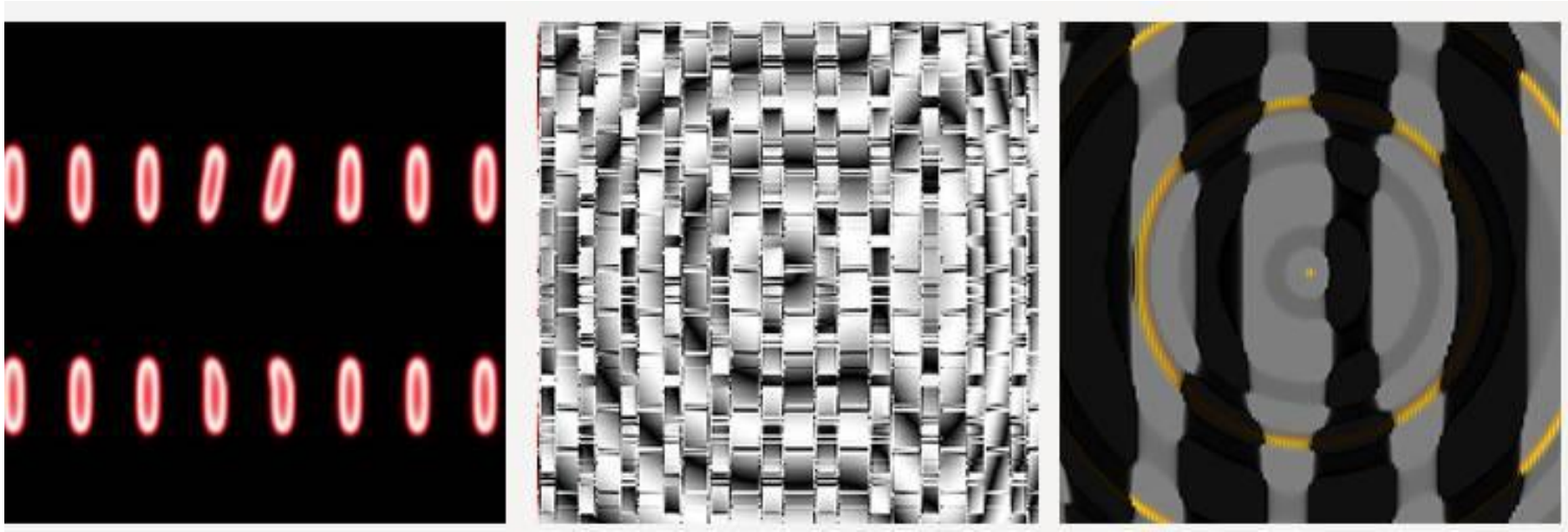
- Regulations changes because of AI solutions
- AI powered Cyber Attacks
- **Attacks against AI solutions**

NOTE: This research has been leaded by AIMS Serbia (Autonomne Inteligentne Mašine i Sistemi)

Evolution Phases of Adversarial Machine Learning

1st generation attacks

1st Generation Attacks: functional but not acceptable by humans



a digital clock | a crossword puzzle | a king penguin

An attempt to confuse ML algorithms, forcing them to make wrong decision
(in this case, false positives), with obviously “wrong data” detectable by humans

NOTE: Prof. Branimir Todorovic's research, AIMS owner, showcased by AST at the 2019 ISC2 Conference in Hong Kong

Evolution Phases of Adversarial Machine Learning

2nd generation of attacks

2nd Generation Attacks: efficient and acceptable by humans

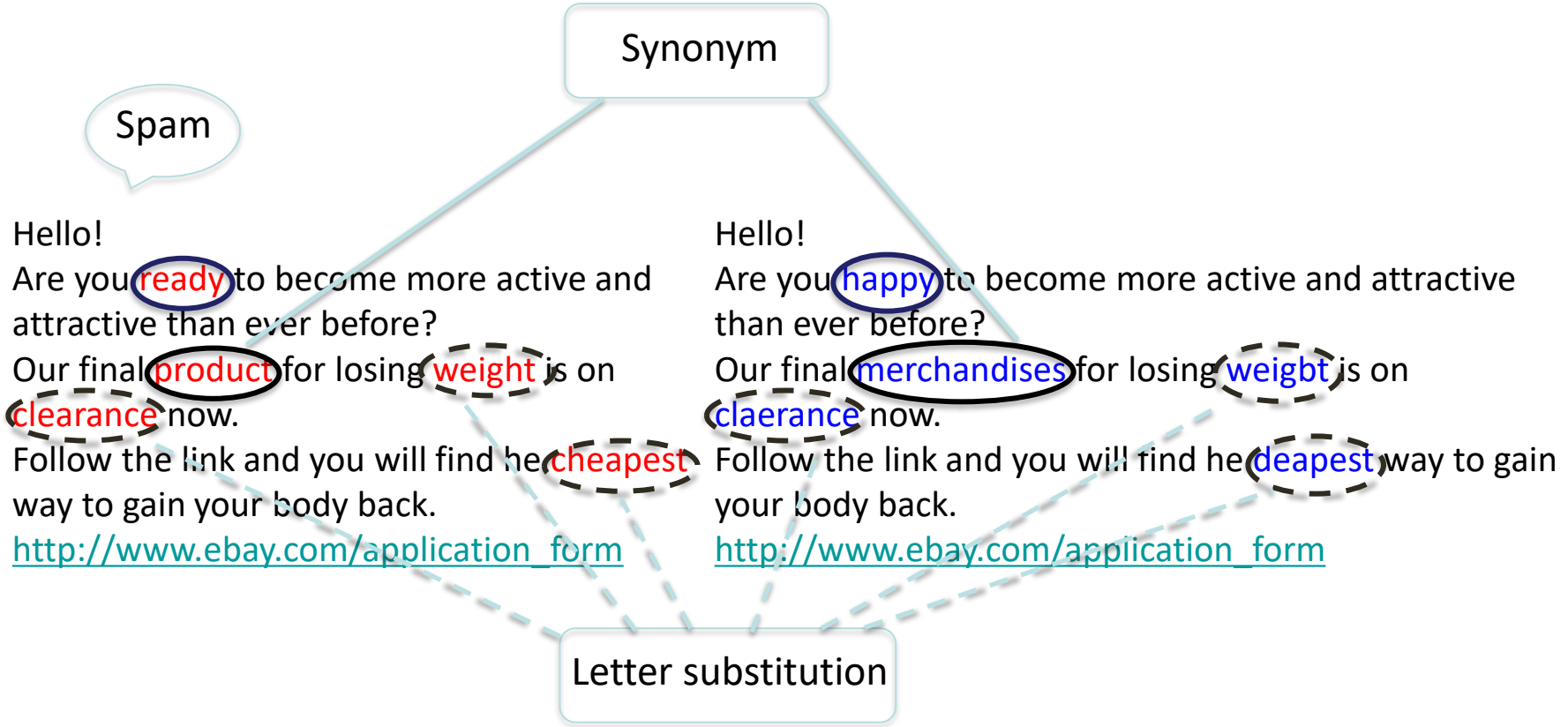
Note: These are real life objects. Their photos were using in adversarial attacks.

An attempt to confuse ML algorithms, forcing them to make wrong decision
(in this case, false negatives), with "wrong data" detectable, but **acceptable** by humans

NOTE: Prof. Branimir Todorovic's research, AIMS owner, showcased by AST at the 2019 ISC2 Conference in Hong Kong

Evolution Phases of Adversarial Machine Learning

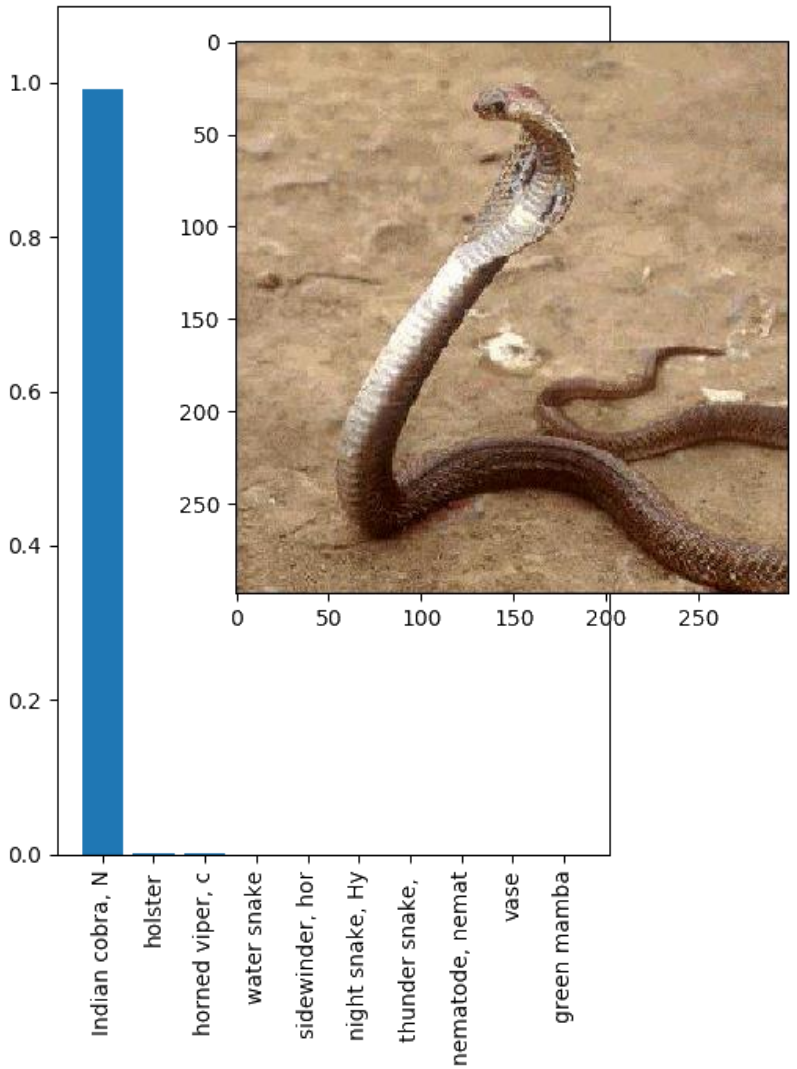
2nd generation of attacks – detectable, but acceptable by humans

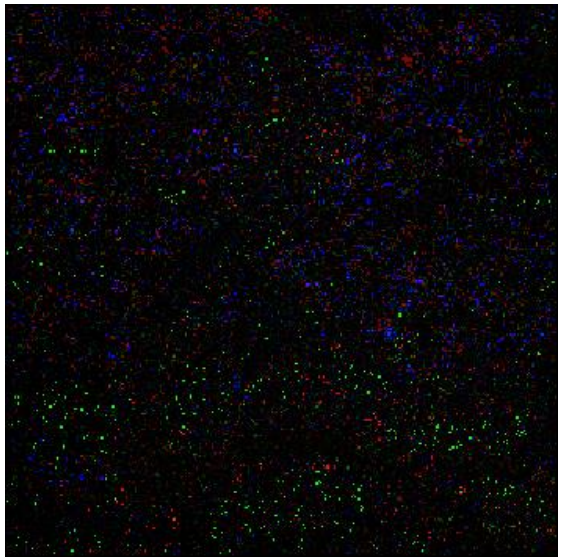


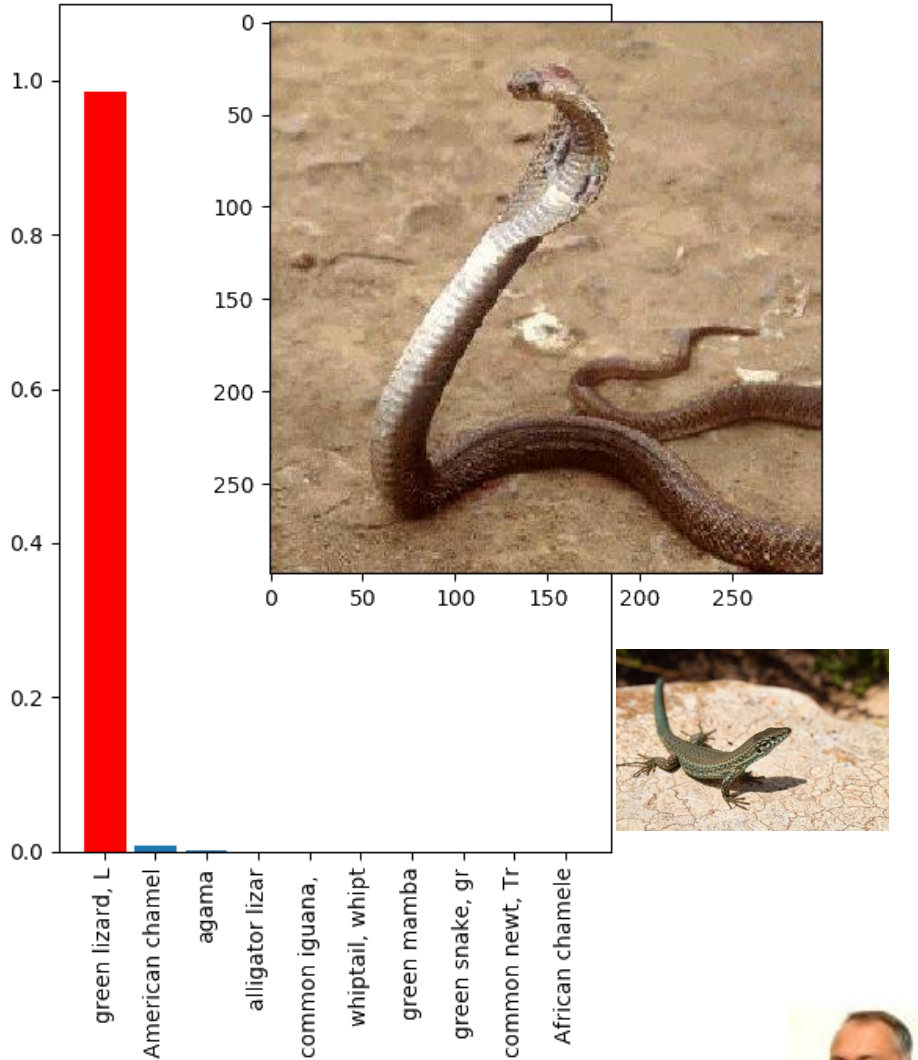
NOTE: Prof. Branimir Todorovic's research, AIMS owner, showcased by AST at the 2019 ISC2 Conference in Hong Kong

Example of ML Attacks Research I

confusing AI algorithm

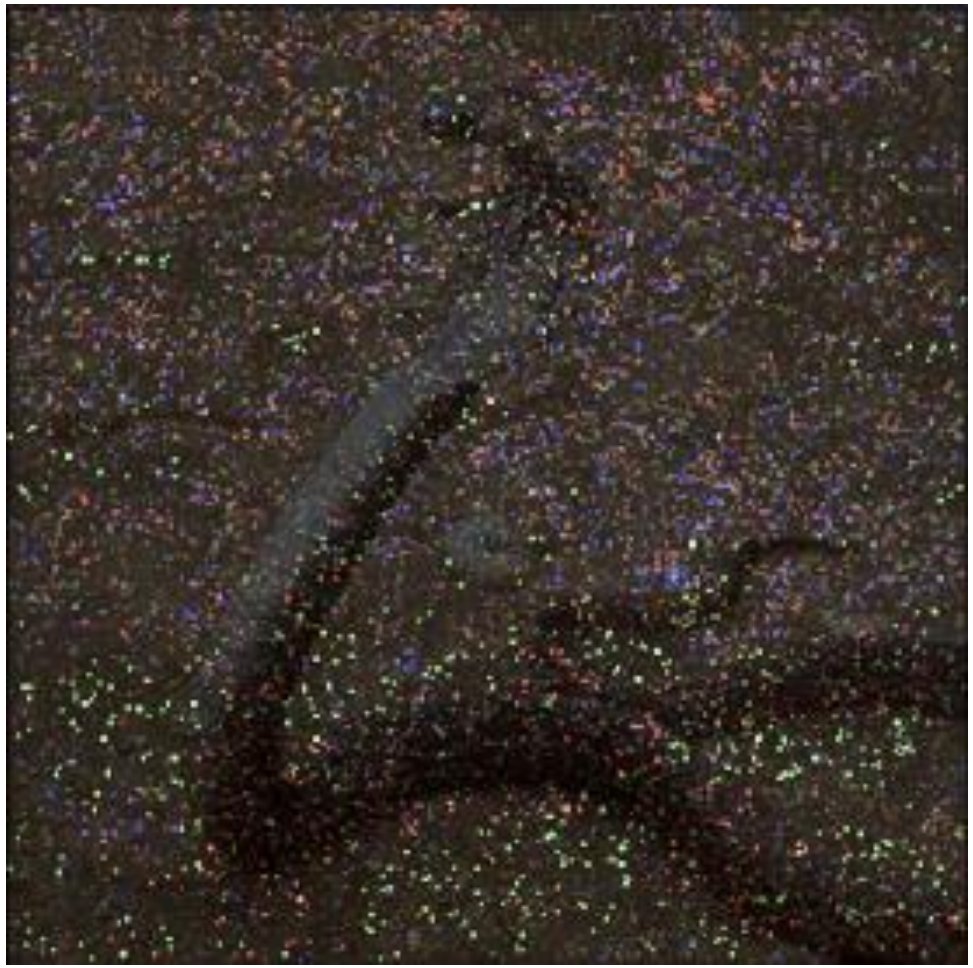






NOTE: Prof. Branimir Todorovic's research, AIMS owner, showcased by AST at the 2019 ISC2 Conference in Hong Kong

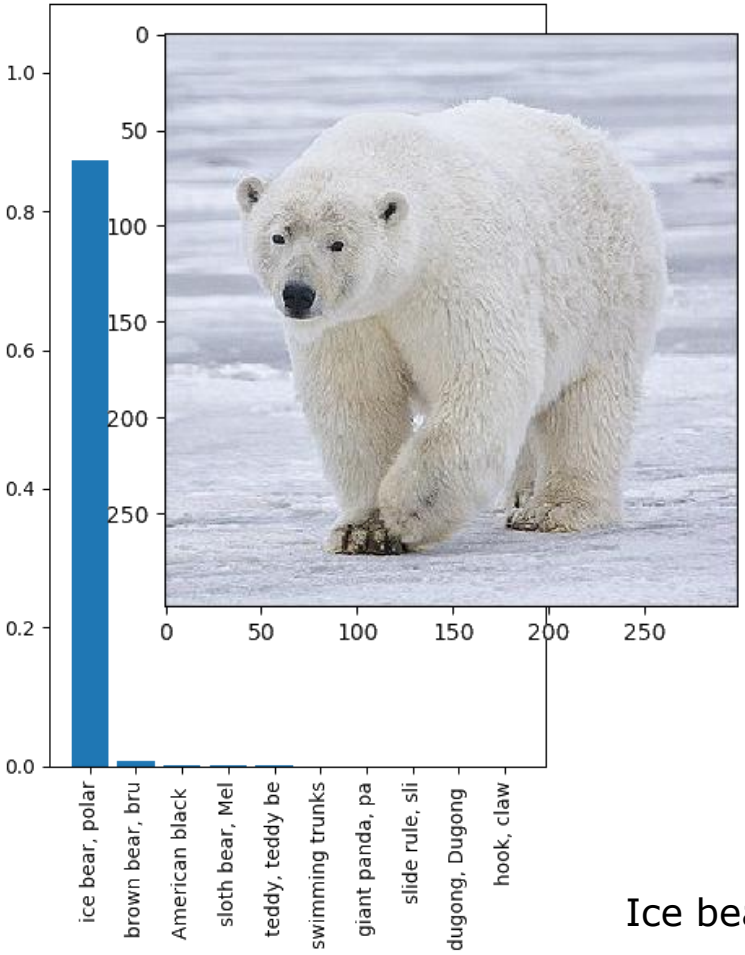
Deep Dive into Changes

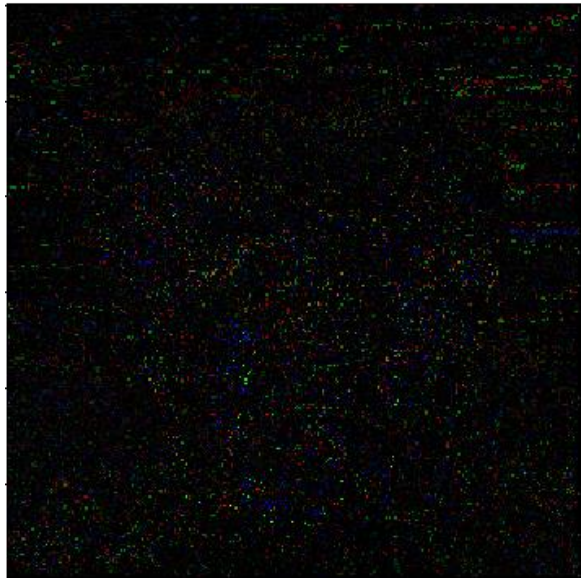


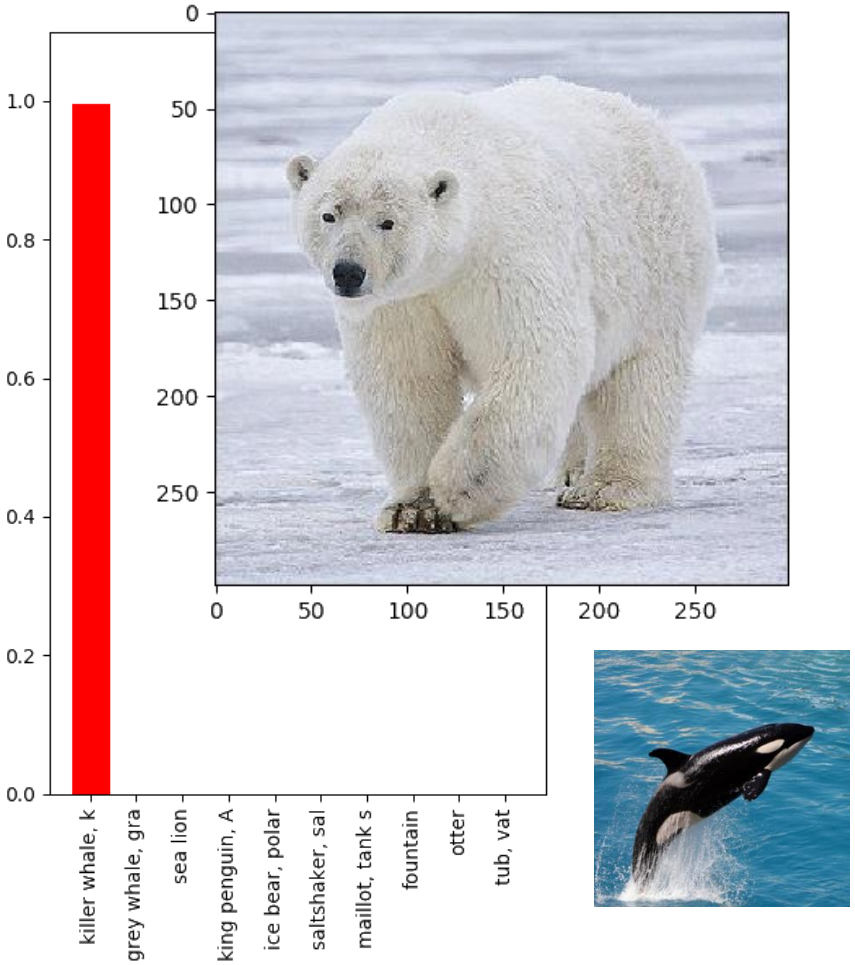
NOTE: Prof. Branimir Todorovic's research, AIMS owner, showcased by AST at the 2019 ISC2 Conference in Hong Kong

Example of ML Attacks Research II

forcing AI algorithm to see what we want it to see



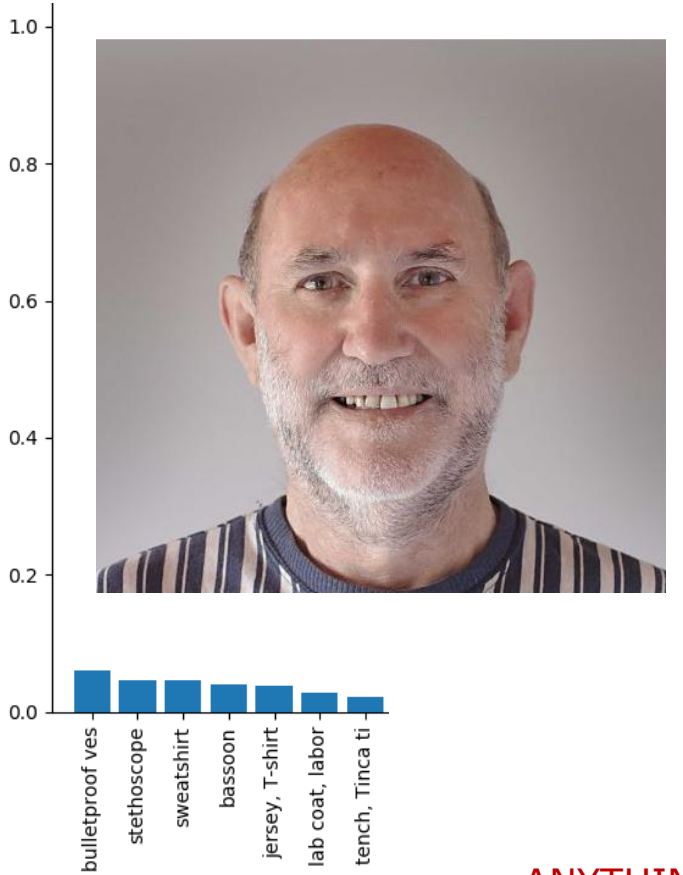


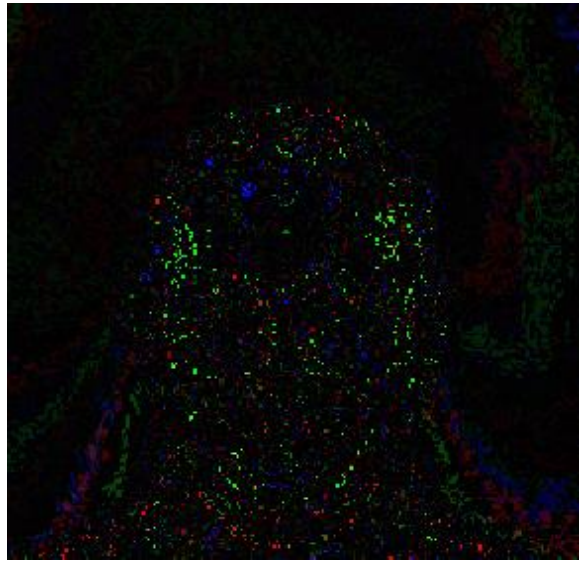


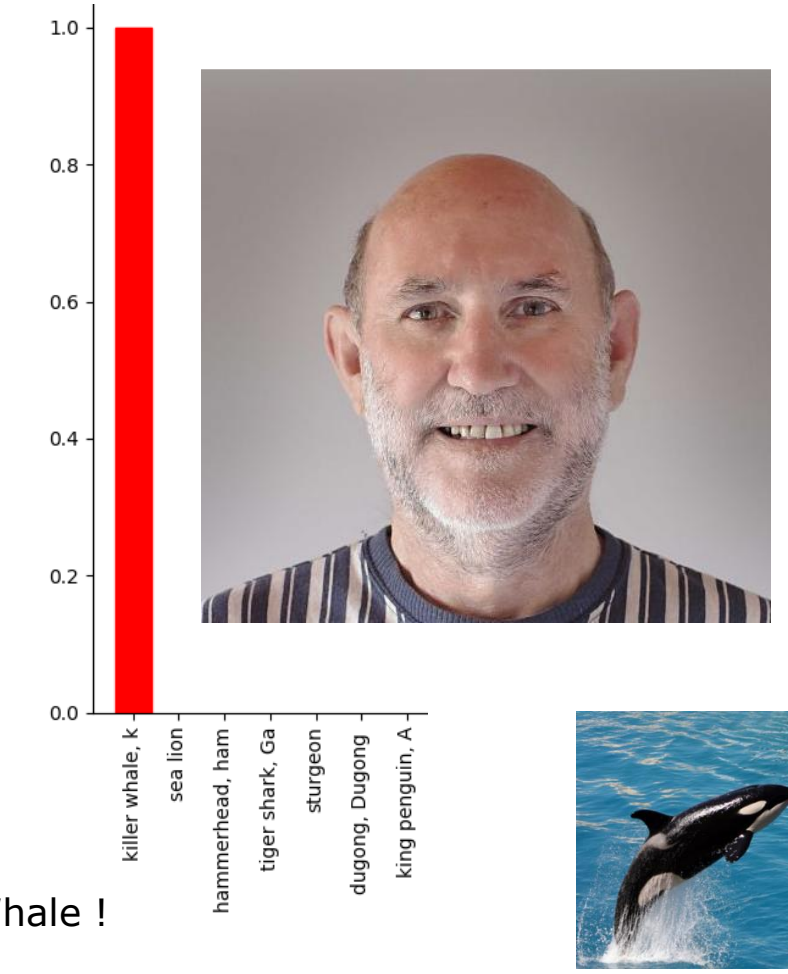
Ice bear + carefully defined noise = Killer Whale !

NOTE: Prof. Branimir Todorovic's research, AIMS owner, showcased by AST at the 2019 ISC2 Conference in Hong Kong

Targeted Attack







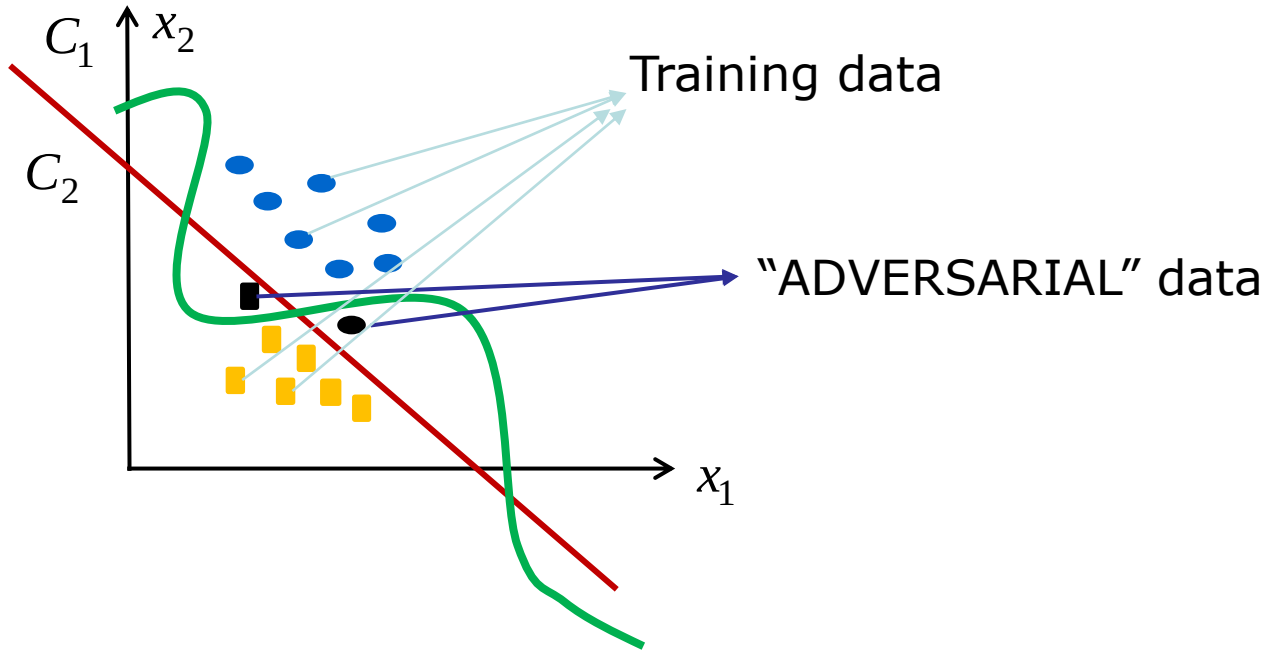
ANYTHING + carefully defined Noise = Killer Whale !

We can calculate the noise for ANY INPUT to force result that we want.

NOTE: Prof. Branimir Todorovic's research, AIMS owner, showcased by AST at the 2019 ISC2 Conference in Hong Kong

How Does It Work?

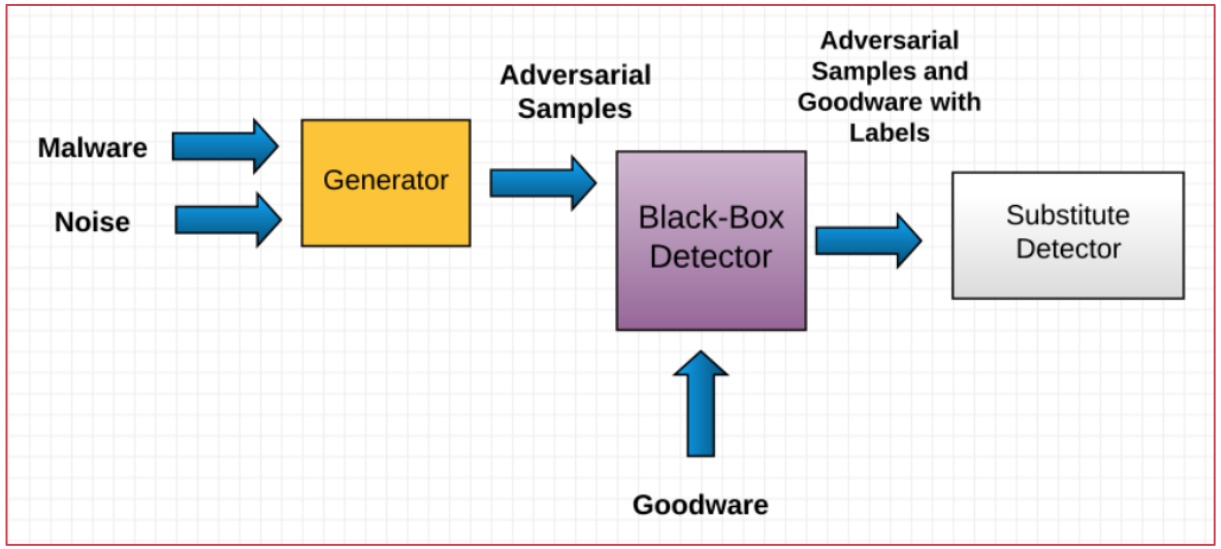
- Often, classifiers are tasked with telling apart "good" from "bad"
  - spam vs. non-spam
  - benign vs. malicious software
  - intrusion detection

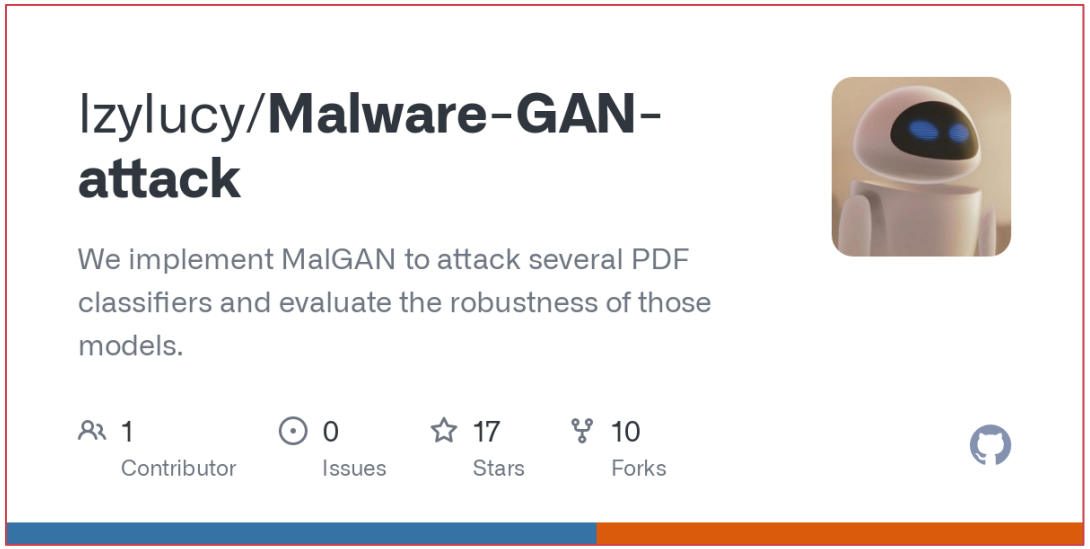


There is no perfect ML algorithm, and there will never be.
We are using data "on the edge" for confusing imperfect ML algorithms.

NOTE: Prof. Branimir Todorovic's research, AIMS owner, showcased by AST at the 2019 ISC2 Conference in Hong Kong

Are there any available automated tools?





Cyber criminals perform attacks against next-generation anti-malware systems, even without knowing the machine learning technique used (black box attacks). One of these techniques is MalGAN

ATLAS™

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below.

& indicates an adaptation from ATT&CK. Click on links to learn more about each item, or view ATLAS tactics and techniques using the links at the top navigation bar.

Reconnaissance &	Resource Development &	Initial Access &	ML Model Access	Execution &	Persistence &	Defense Evasion &	Discovery &	Collection &	ML Attack Staging	Exfiltration &	Impact &
5 techniques	7 techniques	4 techniques	4 techniques	2 techniques	2 techniques	1 technique	3 techniques	3 techniques	4 techniques	2 techniques	7 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution &	Poison Training Data	Evade ML Model	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Adversarial ML Attack Capabilities	Evade ML Model	Physical Environment Access				Discover ML Artifacts	Data from Local System &	Verify Attack		Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access						Craft Adversarial Data		Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets										Cost Harvesting
	Poison Training Data										ML Intellectual Property Theft
	Establish Accounts &										System Misuse for External Effect

AIMS Serbia

Autonomne Inteligentne Mašine i Sistemi

&

AST Serbia

Advanced Security Technologies

Let us know if you would like us to attack YOUR ML solution!

branimirtodorovic.ai [@] gmail.com

v t [@] a s t l t d . c o