



AI and Cybersecurity: A viable partnership?

Sokratis K. Katsikas

Professor, Dept. of Information Security & Communications
Technology, NTNU

Director, SFI Norwegian Center for Cybersecurity in Critical Sectors
(NORCICS), NTNU

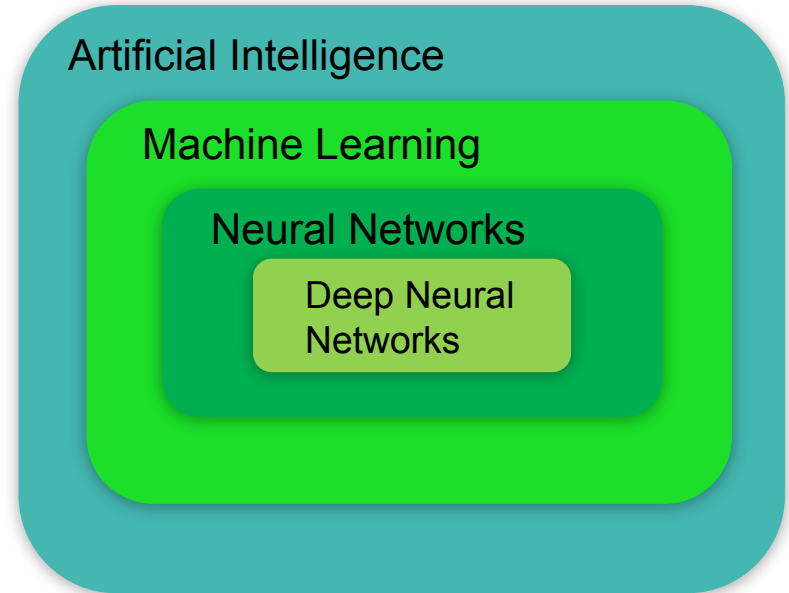
sokratis.katsikas@ntnu.no

Agenda

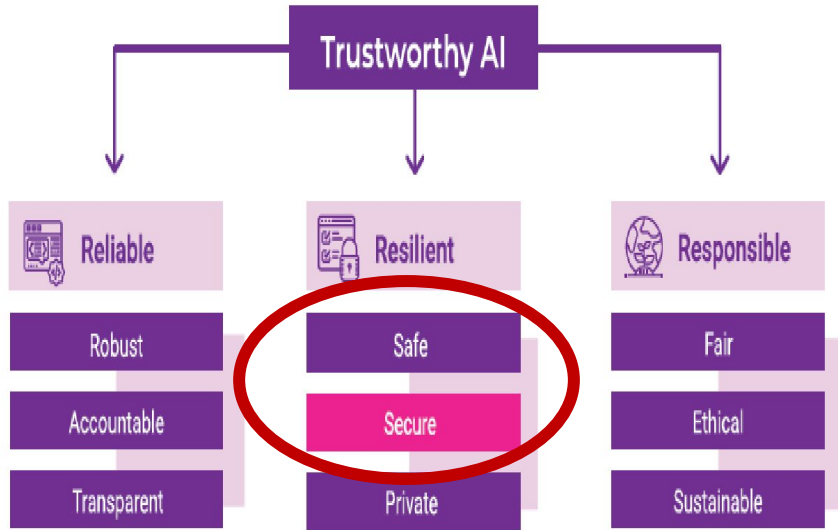
- Introduction
- AI for defensive cybersecurity
- AI for offensive cybersecurity
- Cybersecurity of AI
- Is there anything we can do?
- Take aways

Artificial Intelligence

- Artificial intelligence (AI) has been first defined by John McCarthy in 1955 as “*the science and engineering of making intelligent machines*”
- AI Systems display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals.

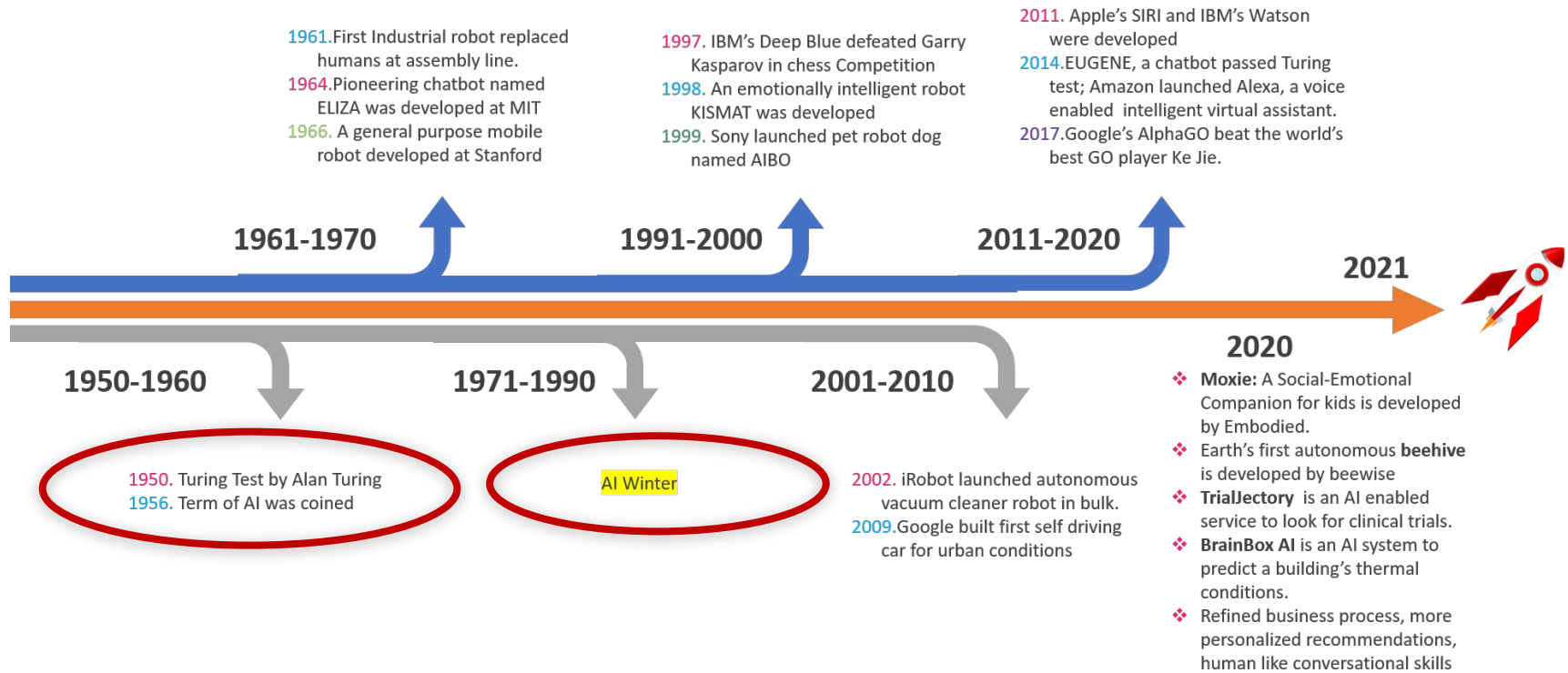


AI systems must be trustworthy



- How do AI and cybersecurity relate?
 - AI is used to improve defensive cybersecurity
 - AI is used to improve offensive cybersecurity
 - Cybersecurity is used to defend AI systems from cyber attacks

ADVERSA, “The road to secure and trusted AI”, v.1.1, 2020



<https://connectjaya.com/ai-timeline/>

Use cases of AI for defensive cybersecurity

- Intrusion detection
- Detection of malicious objects (e.g., documents, websites)
- Defense against Denial-of-Service attacks
- Detection of anomalous or malicious activities through the analysis of network traffic
- Digital forensics support, retrospective incident investigation, attack attribution support
- Identification of fake and compromised Online Social Network accounts
- Identification of fake and/or compromised user/employee accounts
- Threat intelligence acquisition and handling
- Decision support systems and recommenders for SOC, CERT, InfoSec and other security teams
- Authentication (context-based, biometrics)
- Vulnerability Detection
-

Some challenges and future work

- High precision requirements in many cybersecurity applications (e.g., very low FP rates in detection)
- Challenges of practically meaningful evaluation and comparison of ML models (in particular because in cybersecurity ML is often used as part of a much larger technology stack)
- Challenge of designing ML-based IDSs that will perform better than “conventional” techniques under real-world conditions

Is AI always better?



Can Industrial Intrusion Detection Be SIMPLE?

Konrad Wolsing^{1,2(✉)}, Lea Thiemt², Christian van Sloun², Eric Wagner^{1,2}, Klaus Wehrle², and Martin Henze^{1,3}

¹ Cyber Analysis and Defense, Fraunhofer FKIE, Bonn, Germany

{konrad.wolsing,eric.wagner,martin.henze}@kie.fraunhofer.de
² Communication and Distributed Systems, RWTH Aachen University, Aachen, Germany

{wolsing,thiemt,sloun,wagner,wehrle}@comsys.rwth-aachen.de

³ Security and Privacy in Industrial Cooperation, RWTH Aachen University, Aachen, Germany
henze@cs.rwth-aachen.de

Abstract. Cyberattacks against industrial control systems pose a serious risk to the safety of humans and the environment. Industrial intrusion detection systems oppose this threat by continuously monitoring industrial processes and alerting any deviations from learned normal behavior. To this end, various streams of research rely on advanced and complex approaches, i.e., artificial neural networks, thus achieving allegedly high detection rates. However, as we show in an analysis of 70 approaches from related work, their inherent complexity comes with undesired properties. For example, they exhibit incomprehensible alarms and models only specialized personnel can understand, thus limiting their broad applicability in a heterogeneous industrial domain. Consequentially, we ask whether industrial intrusion detection indeed has to be complex or can be SIMPLE instead, i.e., Sufficient to detect most attacks, Independent of hyper-parameters to dial-in, Meaningful in model and alerts, Portable to other industrial domains, Local to a part of the physical process, and computationally Efficient. To answer this question, we propose our design of four SIMPLE industrial intrusion detection systems, such as simple tests for the minima and maxima of process values or the rate at which process values change. Our evaluation of these SIMPLE approaches on four state-of-the-art industrial security datasets reveals that SIMPLE approaches can perform on par with existing complex approaches from related work while simultaneously being comprehensible and easily portable to other scenarios. Thus, it is indeed justified to raise the question of whether industrial intrusion detection needs to be inherently complex.

1 Introduction

Cyberattacks against Industrial Control System (ICSs) with the goal of financial gains, damaging equipment, or even risking human lives by blocking normal operations or injecting false data are becoming more prevalent [7]. Recent examples of such attacks include the attempted poisoning of a Florida city's water supply

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
V. Atthuri et al. (Eds.): ESORICS 2022, LNCS 13656, pp. 674–694, 2022.
https://doi.org/10.1007/978-3-031-17145-7_38

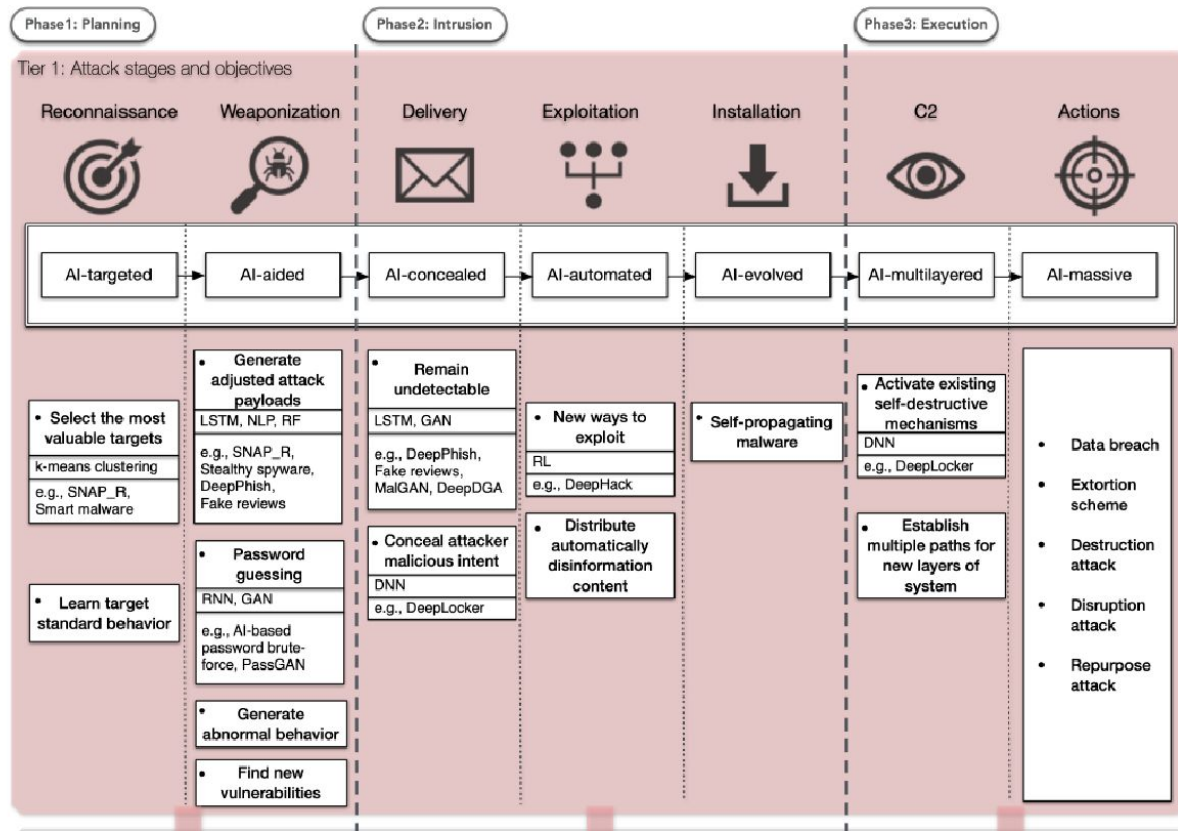
	Detection Method (unique publications)	SWaT (63)	WADI (24)	HAI (6)
Artificial NNs (44)	Autoencoder (15)	[12, 16, 26, 36, 40, 41, 47, 52, 58, 62, 68, 75, 77, 83]	[12, 62]	[45]
	Other NN (12)	[1, 20, 22, 29, 34, 49, 50, 65, 70, 71, 73, 78]	[1, 22, 29, 34, 50, 70, 73]	–
	RNN (9)	[6, 30, 39, 53–56]	[30, 56]	[13, 42]
	GAN (5)	[5, 14, 44, 60, 64]	[14, 60]	–
	DNN (3)	[43, 46, 66]	–	–
Graphs (6)	Automata (3)	[15, 57, 79]	[79]	–
	Other (3)	[17, 35]	[17, 35, 69]	–
Miscellaneous (20)	Invariants (3)	[33, 74, 82]	[33, 82]	–
	Linear algebra (3)	[11, 24, 63]	–	–
	Classifier (2)	[9, 25]	[25]	–
	Fingerprinting (2)	[2, 4]	[2]	–
	Matrix Profiles (2)	[8, 10]	–	–
	Other (8)	[18, 32, 48, 80, 81]	[67, 80, 81]	[21, 48, 61]

- We analyze the current state of IIDS research. Our study of 70 approaches unveils limitations w.r.t. deployability, computational complexity, generalizability, focus on non-stealthy attacks, and incomprehensibility of alarms (Sect. 3).

We then compare the performance of our SIMPLE IIDSs against state-of-the-art complex related work alongside four industrial datasets. Our results show that SIMPLE IIDSs detect more attacks than complex related work detects on average and can be ported effortlessly across industrial domains (Sect. 5).

Improving attackers' capabilities

- AI can make cyber attacks
 - More effective; More targeted; Cheaper; Faster; More autonomous; more difficult to attribute; more interactive
- Use cases
 - Usage of AI for evading authentication controls (e.g., CAPTCHA)
 - Autonomous/automatic movement in cyber attacks
 - Enhanced social engineering and identity theft
 - Massive spear phishing attacks (highly targeted)
 - Highly interactive and scalable social engineering attacks
- Challenges and future work
 - Sufficient mitigation approaches missing
 - Use AI to defend against AI



Cyber Kill Chain

Nektaria Kaloudi and Jingyue Li. 2020. The AI-Based Cyber Threat Landscape: A Survey. ACM Comput. Surv. 53, 1, Article 20 (January 2021), 34 pages. <https://doi.org/10.1145/3372823>



adversarial
perturbation



88% **tabby cat**

99% **guacamole**

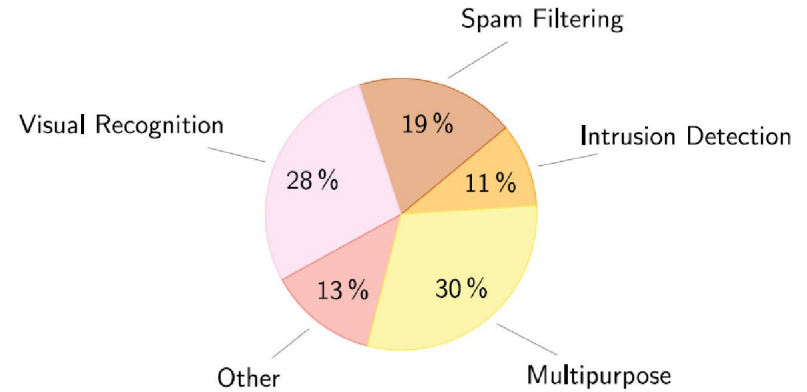
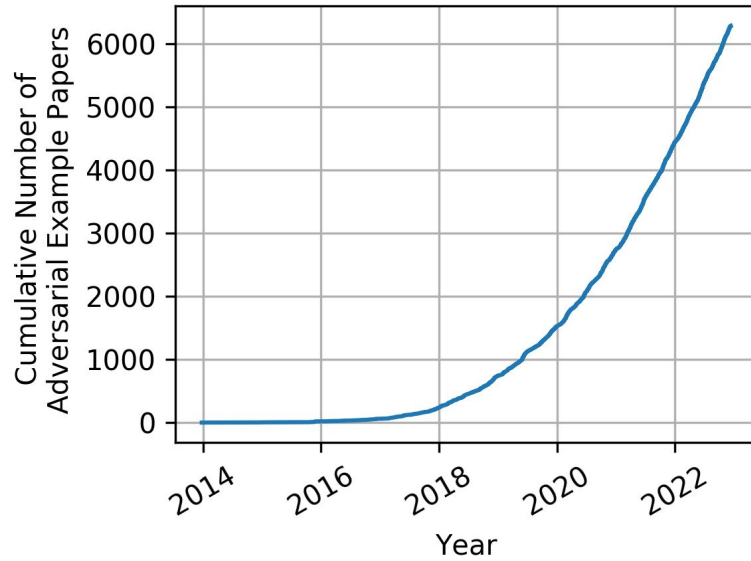
A Simple Explanation for the Existence of
Adversarial Examples
with Small Hamming Distance

Adi Shamir¹, Itay Safran¹, Eyal Ronen², and Orr Dunkelman³

¹ Computer Science Department, The Weizmann Institute, Rehovot, Israel

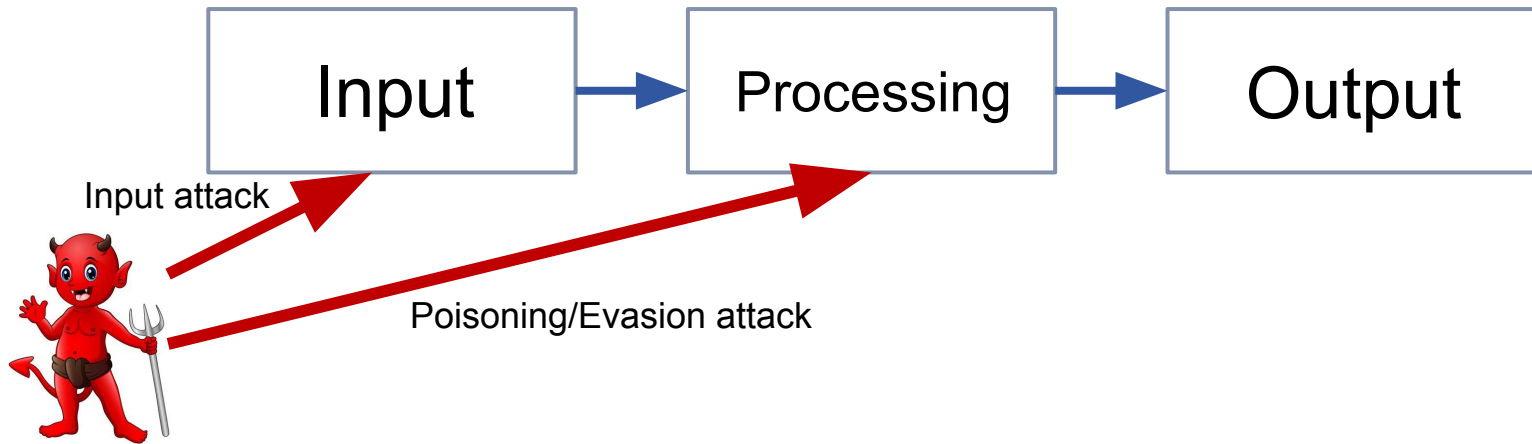
² Computer Science Department, Tel Aviv University, Tel Aviv, Israel

³ Computer Science Department, University of Haifa, Israel



<https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html>

Nikolaos Pitropakis, Emmanouil Panaousis, Thanassis Giannetsos, Eleftherios Anastasiadis, George Loukas, “A taxonomy and survey of attacks against machine learning”, Computer Science Review, Volume 34, 2019, <https://doi.org/10.1016/j.cosrev.2019.100199>



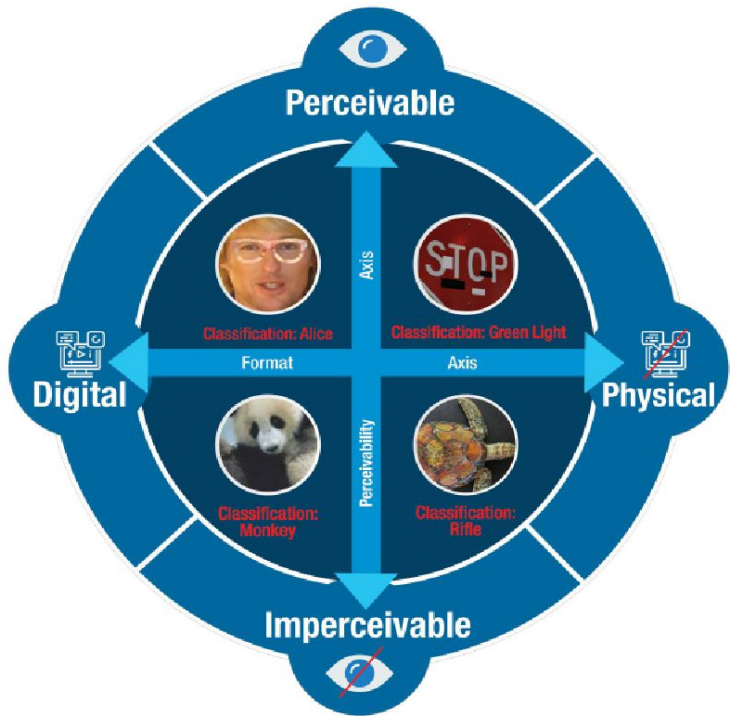
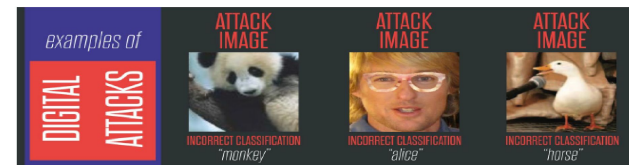
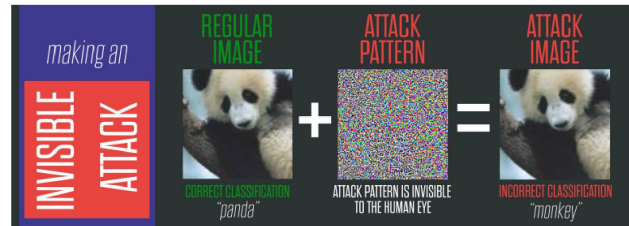
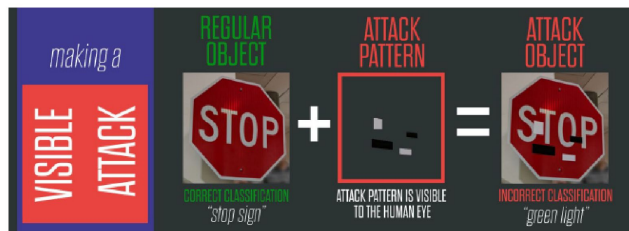
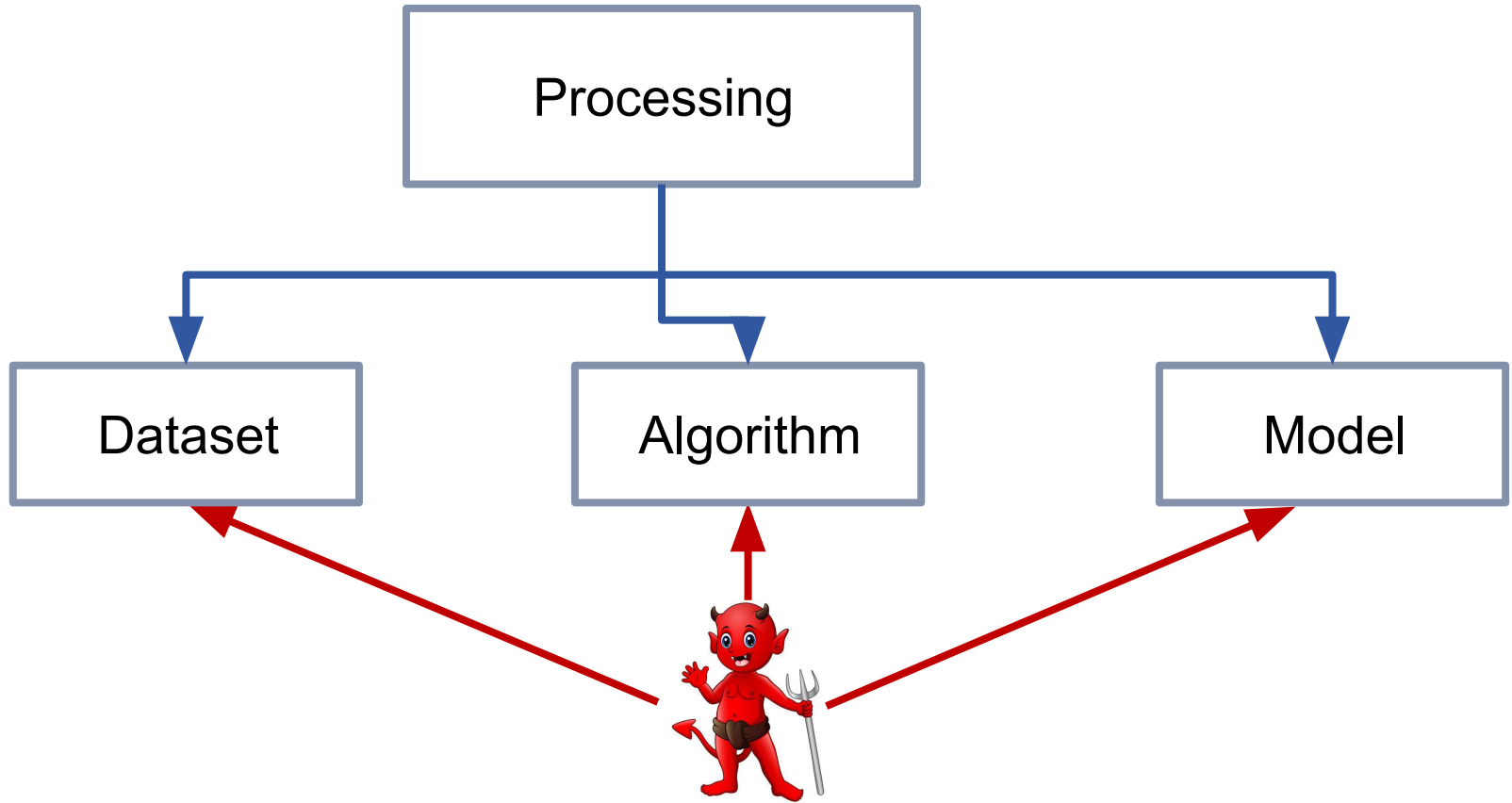
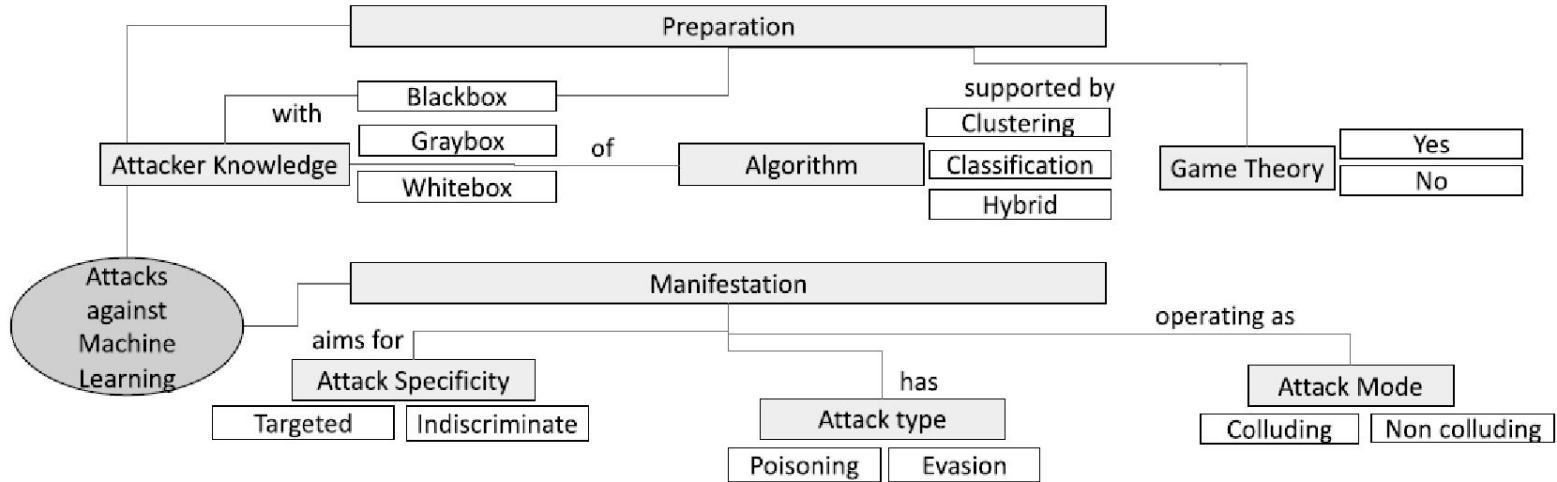
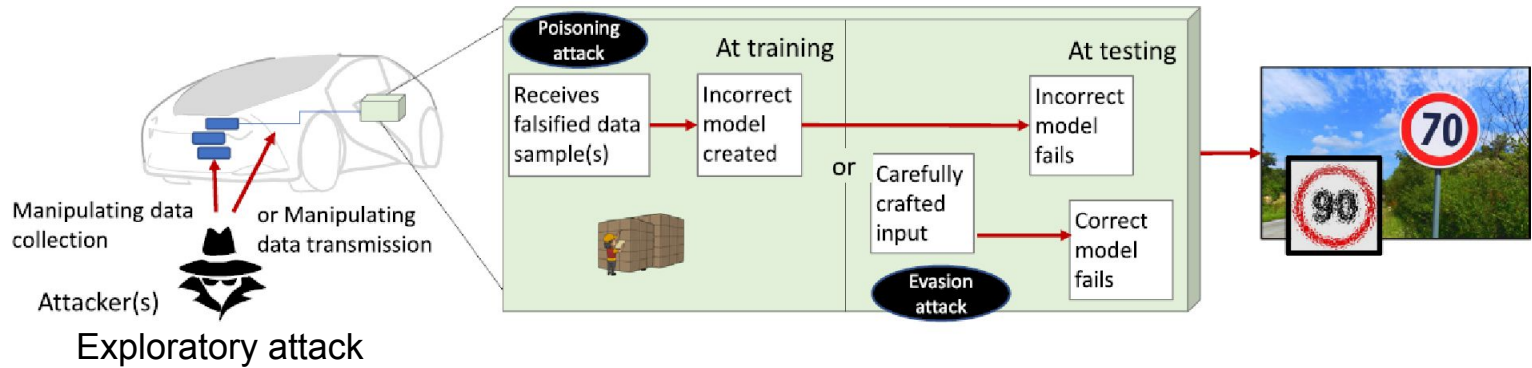


Figure 2: Taxonomy for categorizing input attacks. The horizontal axis characterizes the format of the attack, either in the physical world or digital. The vertical axis characterizes the perceivability of the attack, either perceivable to humans or imperceivable to humans. (See footnote⁸ for thumbnail images citations.)

M. Comier, "Attacking Artificial Intelligence", 2019





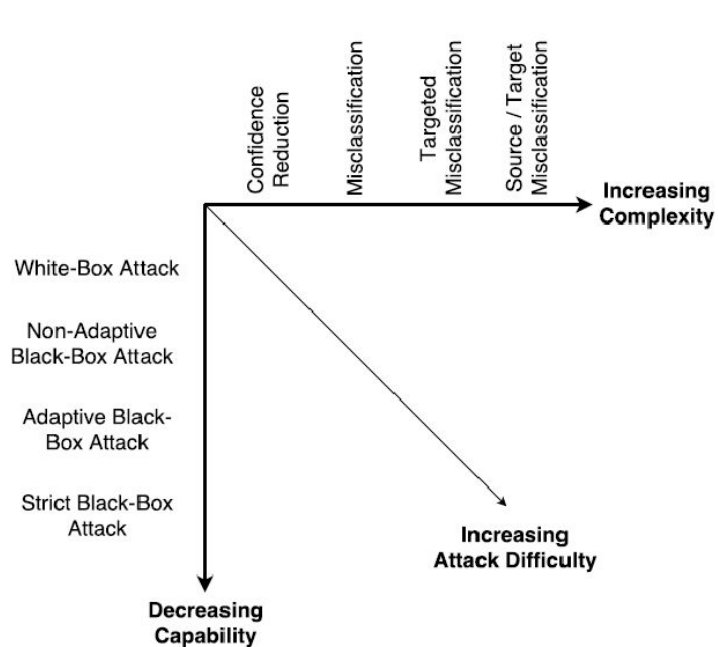


Nikolaos Pitropakis, Emmanouil Panaousis, Thanassis Giannetsos, Eleftherios Anastasiadis, George Loukas, "A taxonomy and survey of attacks against machine learning", Computer Science Review, Volume 34, 2019, <https://doi.org/10.1016/j.cosrev.2019.100199>

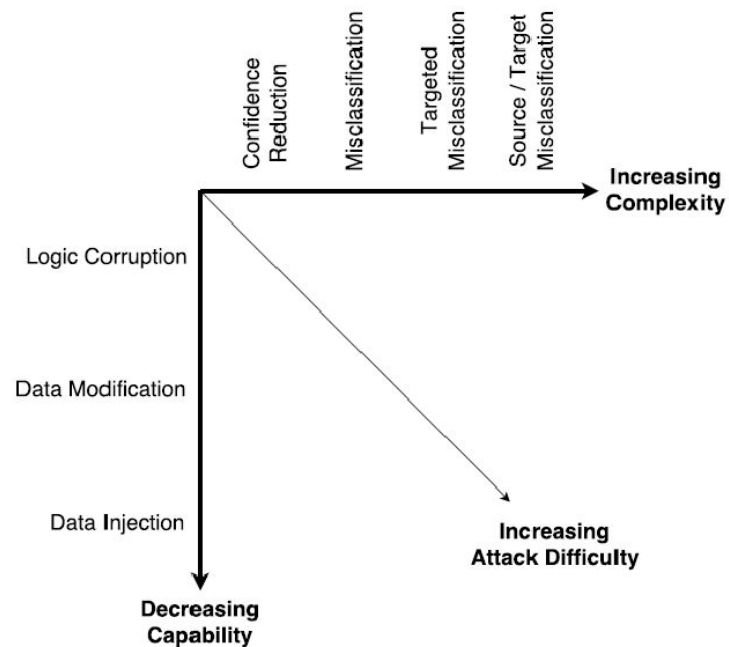
Attacker capabilities and goals

- Training Phase Capabilities
 - Data Injection
 - Data Modification
 - Logic Corruption
- Testing Phase Capabilities
 - Black-Box Attacks
 - Non-Adaptive Black-Box Attack
 - Adaptive Black-Box Attack
 - Strict Black-Box Attack
 - White-Box attacks
- Confidence Reduction
- Misclassification
- Targeted Misclassification
- Source/Target Misclassification

It's not so easy to launch an attack



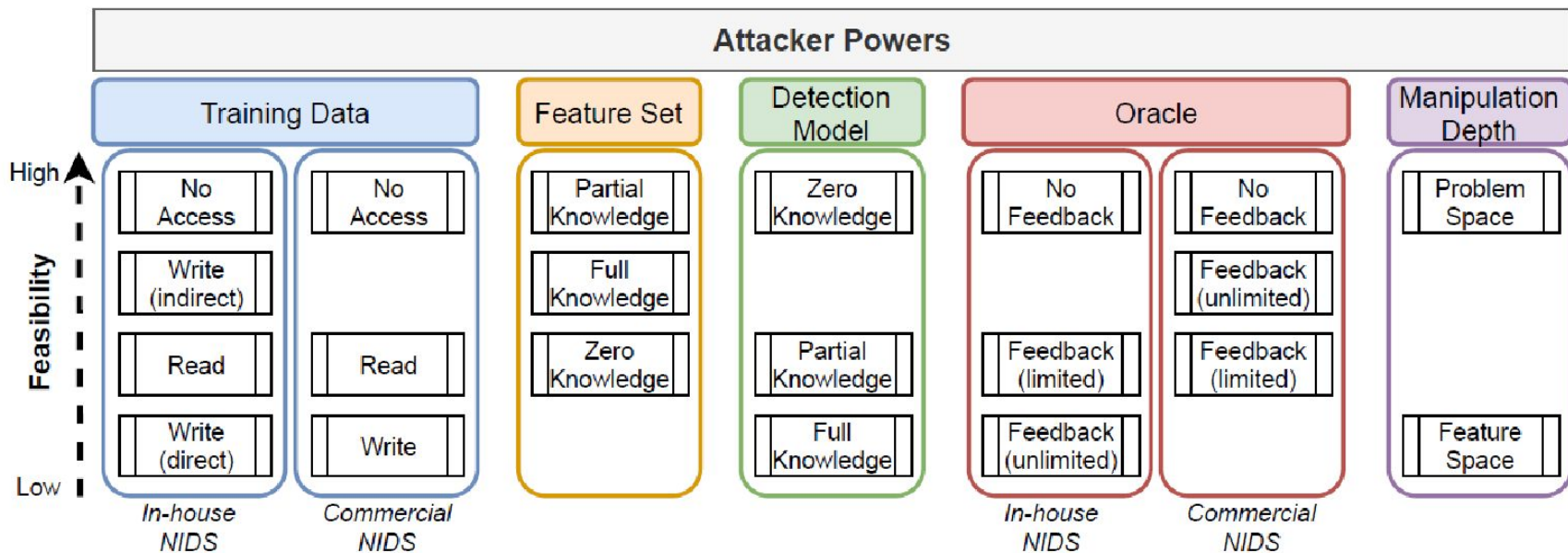
a) Attack Difficulty with respect to adversarial capabilities and goals for Evasion Attacks



b) Attack Difficulty with respect to adversarial capabilities and goals for Poisoning Attacks

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, Debdeep Mukhopadhyay, "Adversarial Attacks and Defences: A Survey", <https://arxiv.org/abs/1810.00069>

When it comes specifically to NIDS



Giovanni Apruzzese, Mauro Andreolini, Luca Ferretti, Mirco Marchetti, and Michele Colajanni. 2022. Modeling Realistic Adversarial Attacks against Network Intrusion Detection Systems. Digital Threats 3, 3, Article 31 (September 2022), 19 pages. <https://doi.org/10.1145/3469659>

Giovanni Apruzzese, Mauro Andreolini, Luca Ferretti, Mirco Marchetti, and Michele Colajanni. 2022. Modeling Realistic Adversarial Attacks against Network Intrusion Detection Systems. Digital Threats 3, 3, Article 31 (September 2022), 19 pages. <https://doi.org/10.1145/3469659>

Paper	Year	Power					Constraints Verified?	Type	Dataset
		Training Data	Feature set	Detection model	Oracle	Manipulation depth			
Kloft et al. [47]	2010	W	—	Full	—	feature	X	poisoning	KDD99
Biggio et al. [17]	2013	—	Full	Partial	—	feature	X	inference	custom
Sethi et al. [79]	2018	—	Full	—	—	feature	✓	inference	KDD99
Warzynski et al. [98]	2018	—	Full	Full	—	feature	X	inference	KDD99
Apruzzese et al. [10, 13]	2018	—	Partial	—	—	feature	✓	inference	CTU13, Botnet2014 IDS2017, CICIDS2018
Lin et al. [57]	2018	—	—	—	∞	feature	✓	inference	KDD99
Li et al. [54]	2018	R, W	Partial	—	—	feature	X	poisoning	KDD99, Kyoto[87]
Wang et al. [97]	2018	—	Full	Full	—	feature	X	inference	KDD99
Yang et al. [102]	2018	R	Partial	—	∞	feature	✓	inference	KDD99
Marino et al. [60]	2018	—	Full	Full	—	feature	X	inference	KDD99
Apruzzese et al. [12]	2019	R, W	Partial	—	—	feature	X	poisoning	CTU13
Hashemi et al. [34, 35]	2019	—	Full	Full	—	feature	X	inference	IDS2017
Usama et al. [93]	2019	—	—	—	∞	feature	✓	inference	KDD99
Khamis et al. [3]	2019	—	Full	Full	—	feature	X	inference	UNSW-NB15
Clemens et al. [23]	2019	—	Full	Full	—	feature	X	inference	Kitsume
Ibitoye et al. [39]	2019	—	Full	Full	—	feature	X	inference	Bot-IoT[48]
Aiken et al. [4]	2019	—	Partial	—	—	problem	✓	inference	IDS2017
Yan et al. [101]	2019	—	—	—	∞	feature	✓	inference	KDD99, IDS2017
Martins et al. [61]	2019	—	Full	Full	—	feature	X	inference	KDD99, IDS2017
Usama et al. [94]	2019	—	—	—	∞	problem	✓	inference	Tor-nonTor[52]
Peng et al. [71]	2019	—	—	—	∞	feature	✓	inference	KDD99, IDS2017
Kuppa et al. [51]	2019	—	Partial	—	100s	feature	✓	inference	CICIDS2018
Wu et al. [99]	2019	—	Partial	—	10s	problem	✓	inference	CTU13
Ayub et al. [15]	2020	—	Full	Full	—	feature	X	inference	IDS2017, TRAbID2017[96]
Novo et al. [65]	2020	—	Full	Full	—	feature	✓	inference	MTA [2]
Chemikova et al. [22]	2020	—	Partial	—	∞	feature	✓	inference	CTU13
Sadeghzadeh et al. [78]	2020	—	Full	Full	—	problem	✓	inference	ICSX2016
Chemikova et al. [22]	2020	—	Full	Full	—	feature	✓	inference	CTU13
Piplai et al. [74]	2020	—	Full	Full	—	feature	X	inference	BigDataCup2019[43]
Alhajjar et al. [6]	2020	—	Full	Full	—	feature	X	inference	KDD99, UNSW-NB15
Apruzzese et al. [9]	2020	—	Partial	None	20s	feature	X	inference	CTU13, Botnet2014
Han et al. [33]	2020	—	Partial	—	100s	problem	✓	inference	Kitsume, IDS2017
Pawlicki et al. [70]	2020	—	Full	Full	—	feature	X	inference	IDS2017
Shu et al. [85]	2020	—	Full	Partial	50s	feature	X	inference	IDS2017
Papadopoulos et al. [67]	2021	R, W	Full	—	—	feature	X	poisoning	Bot-IoT
Pacheco et al. [66]	2021	—	Full	Full	—	feature	X	inference	UNSW-NB15
Anthi et al. [8]	2021	R	Full	—	—	feature	X	inference	ICSX2016
Papadopoulos et al. [67]	2021	—	Full	Full	—	feature	X	inference	Bot-IoT

Challenges and future work

- ML works by learning patterns liable to break easily
- Dependence on data provides a main channel to corrupt a machine learning model
- The “black box” nature of the algorithms makes auditing them difficult
- Attacks are not due to vulnerabilities, but to deep-seated, structural issues at the heart of current ML
- Adversarial attacks represent a threat affecting the reliability of any cyber defense relying on artificial intelligence.

Defensive strategies (examples)

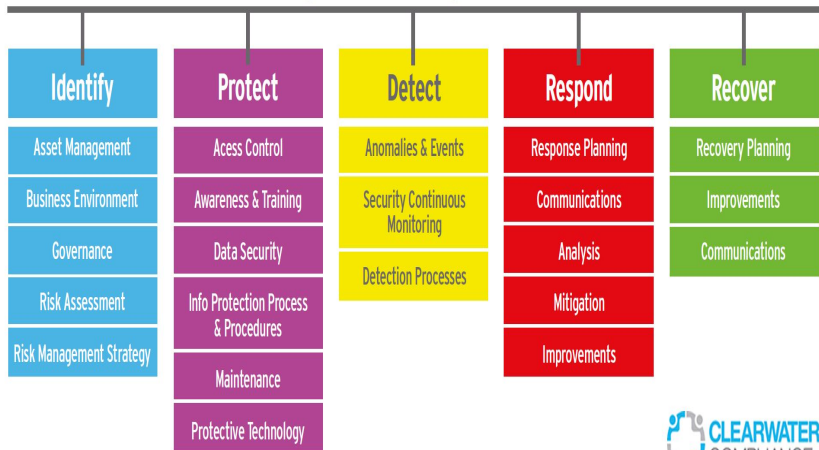
- Adversarial Training
- Gradient Hiding
- Defensive Distillation
- Feature Squeezing
- Blocking the Transferability
- Defense-GAN
- MagNet
- Using High-Level Representation Guided Denoiser
- Using Basis Function Transformations

Several defensive strategies exist, but...

- Adversarial examples are hard to defend because:
 - A theoretical model of the adversarial example crafting process is very difficult to construct
 - Machine Learning models are required to provide proper outputs for every possible input
- Most of the current defense strategies are not adaptive to all types of adversarial attacks, as one method may block one kind of attack but leaves another vulnerability open to an attacker who knows the underlying defense mechanism.
- Implementation of such defense strategies may incur performance overhead and can also degrade the prediction accuracy of the actual model.

A way forward

NIST Cyber Security Framework



A structured, holistic approach is needed





NIST AI 100-1



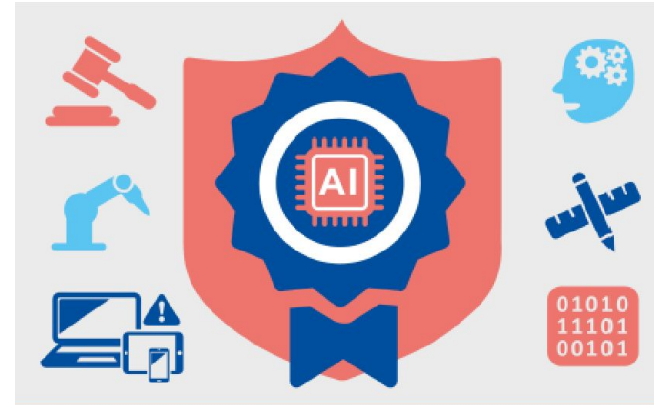
Artificial Intelligence Risk Management Framework (AI RMF 1.0)

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

- Standardization bodies doing relevant work
 - CEN-CENEL
 - EC
 - ETSI
 - ISO-IEC
- NIST



EUROPEAN UNION AGENCY
FOR CYBERSECURITY



CYBERSECURITY OF AI AND STANDARDISATION

MARCH 2023



GREEN AI (SINTEF)



ADAPTIVE AI (IMT)



HUMAN-CENTRIC AI (TUE)



TRUSTWORTHY AI (NTNU)



ENERGY
(INESC)



HEALTHCARE
(ICCS)



MANUFACTURING
(POLIMI)



SPACE
(ECOE)

Available Resources

- Research labs
- Data
- Infrastructure
- DIHIWARE Platform

Strategic Initiatives

- Policy initiatives
- Networks and alliances
- National and international R&I projects

Innovation Paradigm

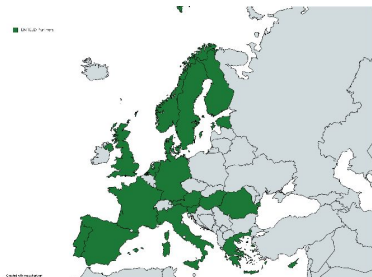
- DIHs and Incubators
- Industry collaborations in ENFIELD and beyond
- Open Cascading Calls: Innovation Scheme

Common Research Vision and Roadmap

Safety and Security Risks Assessment Framework Regulations

Standards and certification

Collaboration, networking and exchange programmes



- A unique **European Centre of Excellence**
- **Fundamental research** in the scientific pillars of **Adaptive, Green, Human-Centric, and Trustworthy AI** within verticals of **healthcare, energy, manufacturing and space**
- **30 consortium members from 18 countries**, including **top-level education and research organisations, large scale businesses, SMEs, and public sector representatives.**
- High impact outputs such as **>75 unique AI solutions (algorithms, methods, simulations, services, data sets and prototypes), 180 scientific high-impact publications and 200 peer-reviewed presentations, four strategic documents.**
- **Exchange and innovation schemes** planned in Open Calls, which will grant **financial support to >76 individual researchers and 18 small-scale projects, education, and training activities such as summer schools and hackathons.**

ENFIELD: European Lighthouse to Manifest Trustworthy and Green AI

Type of Action: HORIZON-RIA

Duration: 36 months (01.09.2023 – 31.08.2026)

Maximum EU contribution: 11,487,793.7 EURO

Coordinator: NTNU

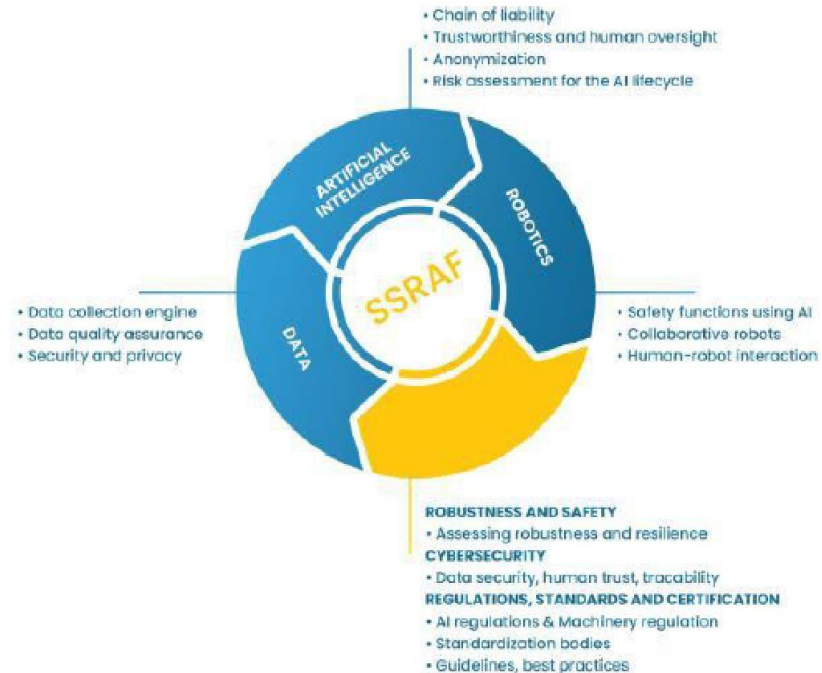
ENFIELD will deliver

1. New risk-aware AI tools that can be trusted to not fall into severe traps, based on autonomous risk-assessment;
2. Benchmarks of experiments for verifying the trustworthiness of some state-of-the-art DNNs on typical safety-critical tasks;
3. Augmented data sets in public domain; DNN learning methods working directly on encrypted data;
4. Automated input or training data anonymization solutions;
5. Design and development of privacy-preserving AI solutions;
6. Risk management framework for AI systems;
7. Data sharing protocols;
8. Methods for testing the security of AI systems.

Dynamic SSRA Framework

- AI Act: risk-based approach
- Different lifecycle of AI-based systems
- Safety and security
- Dynamic nature of AI-based systems
- SSRAF main areas
 - Risk Assessment
 - Quantification of reliability and robustness
 - Functional testing

Figure 1.3 Dynamic SSRAF Framework



NORCICS SFI <https://www.ntnu.edu/norcics>

- Funding for 5(+3) years
- Total budget: 215,645,000 NOK
- Funding: 96,000,000 NOK NFR (41.9%)
- Coordinator (NTNU) + 18 partners (4 research, 14 user)
- Sectors represented: Energy, Manufacturing, Oil & Gas, Security, Healthcare, Police, Process industry



Take aways

- AI systems are being increasingly used in everyday life, including in mission critical systems
- AI systems can improve cybersecurity
- AI systems can be misused
- AI systems are vulnerable to cyberattacks whose impact can be catastrophic; need for resilience
- Therefore, we need:
 - to improve our understanding of the interconnections between the two domains;
 - to leverage this understanding to develop methods and tools to address efficiently and effectively the identified challenges; and
 - to validate the methods and tools to be developed in a number of use cases.

C.S. TECHNOPOLY

OUR MISSION

Making cyber security strategy interesting and fun

- NTNU is building a serious game community
- Devoted to providing society with a deeper understanding of cyber security



A role-playing and scenario game where to win, you will need to learn:

Threat Intelligence



Cybersecurity investments



Security value chain



Systems thinking & Socio-technical transitions



REGISTER HERE



WHEN: WEDNESDAY, 30.AUGUST, kl. 9-12

WHERE: FESTSALEN, KULTURHUSET BANKEN, LILLEHAMMER



Thank you!

“Collaboration = innovation”