ETSI The Standards People

Security Conference 2024

Adversarial AI in ICT

piotr.zuraniewski@tno.nl (ETSI SAI delegate)

TNO innovation for life

14/10/2024

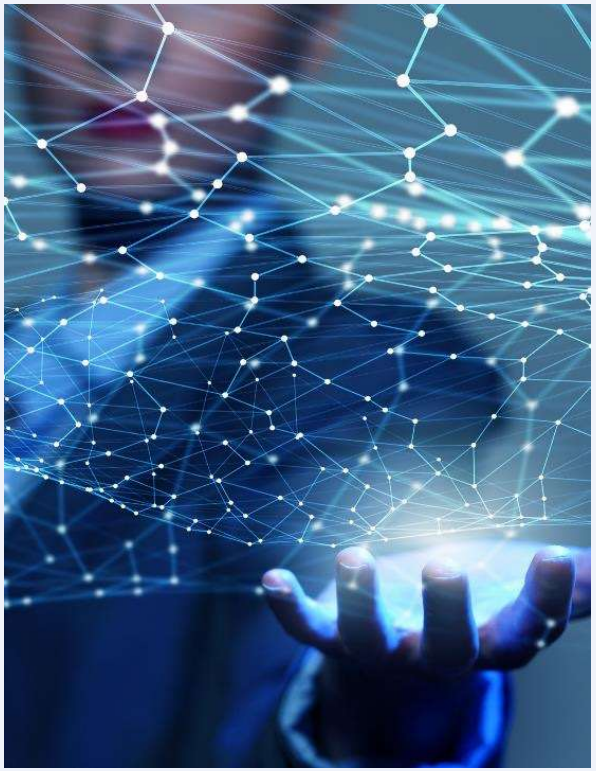
TNO: Netherlands Organisation for Applied Scientific Research

Connecting people and knowledge

Creating innovations

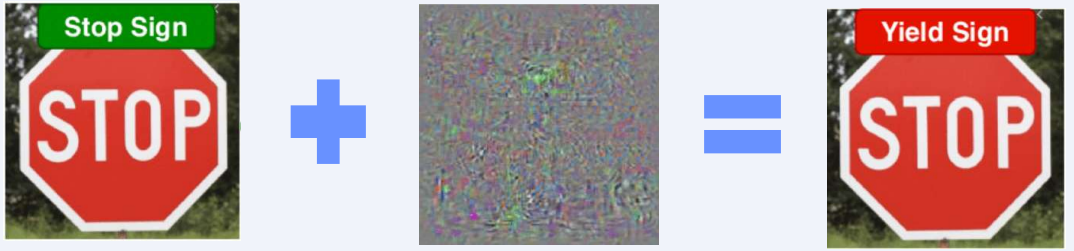
Sustainably strengthening business competitiveness

Sustainably strengthening well-being across society



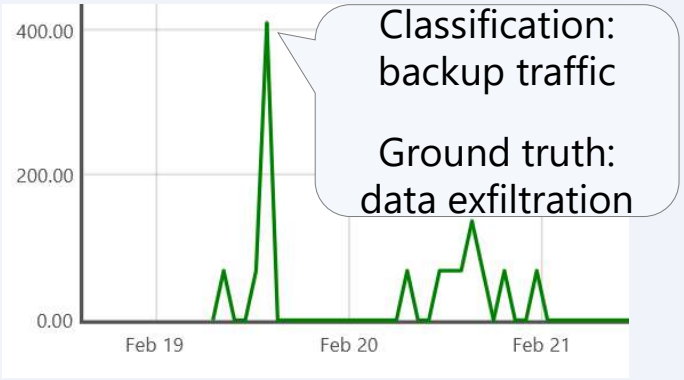
AI poses new type of security problem in ICT

- Situation
  - **AI (to be) used everywhere**
  - E.g., image recognition in self driving cars;
    advanced network management



Authentic Input | Adversarial Perturbation | Adversarial Input

- Complication
  - **AI systems can become targets on their own**
  - E.g., wrongly recognized object; cyberattack misclassified as benign traffic



- **No systematic solution as of now**
  - Most research in image recognition, little in ICT/cloud use cases

NATO OTAN
NATO Communications and Information Agency
NCI AGENCY

TNO innovation for life

TNO projects related to AI security in ICT context

- ADVICE – Adversarial AI in ICT
  - Jointly with NCIA as strategic partner, part of TNO appl.ai multi-year programme
  - Identification of adversarial scenarios in various phases of ICT infra lifecycle

- FNS – Future Network Services
  - TNO leads ecosystem of 60 partners working on 6G, sponsored by Dutch govt.
  - AI-based generation of 6G configs and software as one of the tasks
  - Reach-out to standardization

- Red Teaming AI
  - Internal TNO knowledge building project
  - Ethically attacking your own AI systems to find vulnerabilities

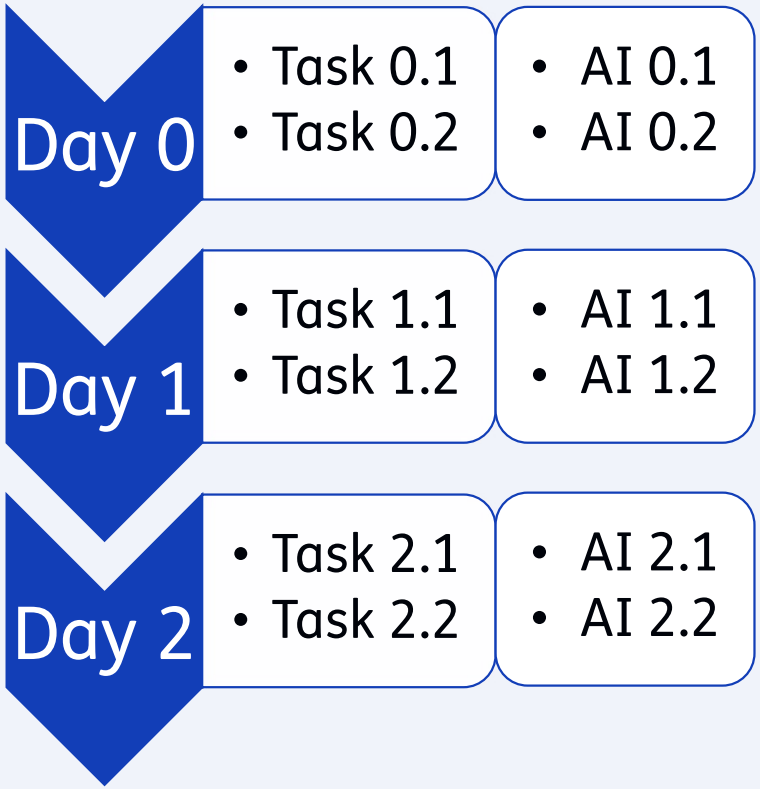
NATO OTAN
NATO Communications and Information Agency
NCI AGENCY

FUTURE 6G NETWORK SERVICES

Rijksoverheid

Adversarial scenarios in ICT infrastructure lifecycle

- We model lifecycle using “Days” structure (see [ETSI_NFV022],[ETSI_OSM])
  - Day 0: Design and plan
  - Day 1: Deployment
  - Day 2: Operations and maintenance
- For each Day, enumerate specific Tasks
  - Example: Day 0, Task 1: Understand strategic/business goals that the ICT system will fulfill
- For each Task, identify AI technique that can be used to fulfill it
  - Example (cnt’d): use LLM as idea generator/sparring partner

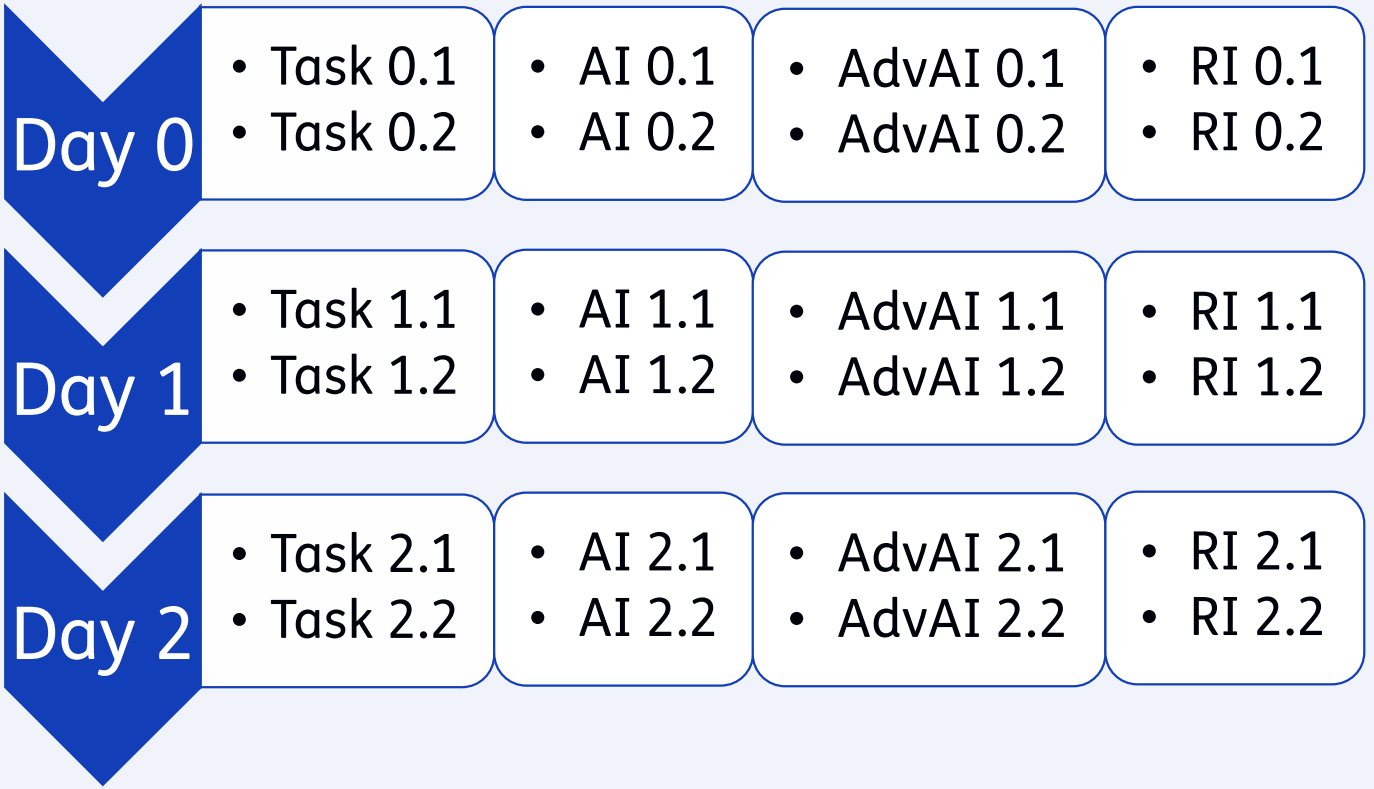


[ETSI_NFV022] ETSI GR NFV-EVE 022 „Network Functions Virtualisation (NFV) Release 5; Architectural Framework; Report on VNF configuration”

[ETSI_OSM] https://osm.etsi.org/docs/user-guide/latest/02-osm-architecture-and-functions.html

Adversarial scenarios in ICT infrastructure lifecycle

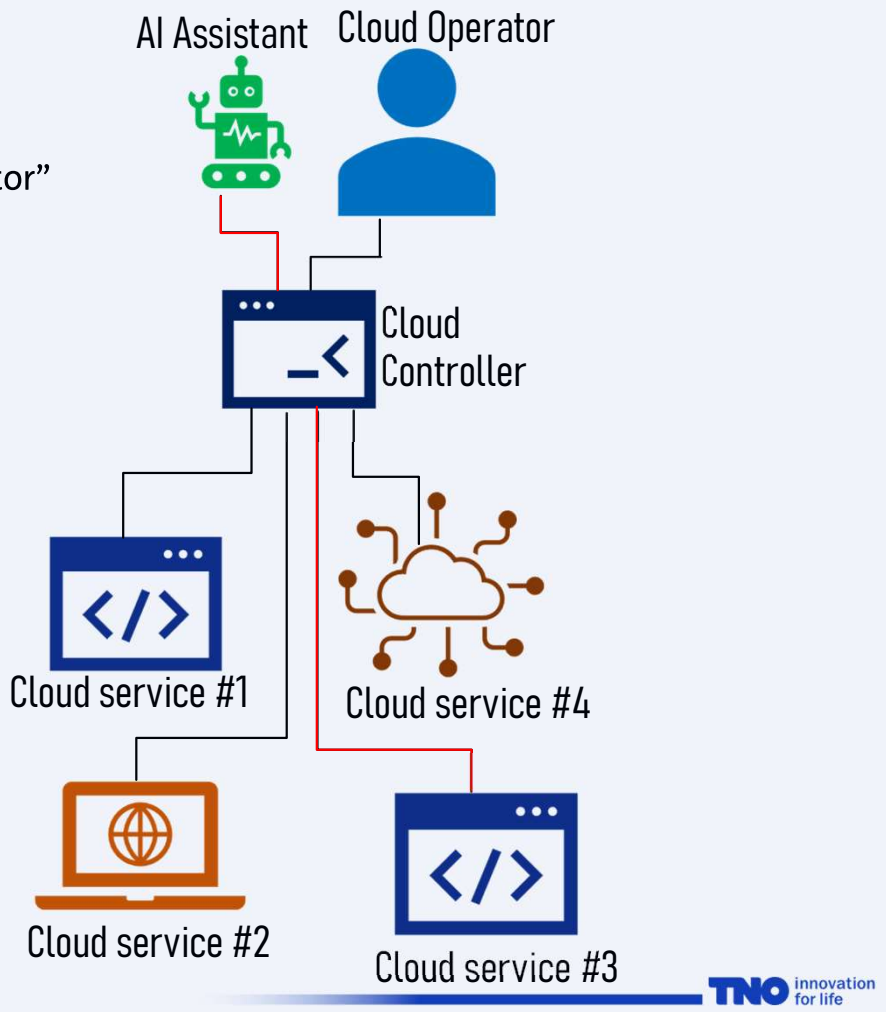
- Next, for each AI technique, identify **adversarial** AI technique
  - Example (cont'd):
    - AI: use LLM
    - **AdvAI: Data extraction**
- Attempt to assess 'risk index' (RI) using e.g., CVSS4
- Consider mitigation measures, both:
  - 'classic' e.g., access control
  - AI-specific e.g., prompt sanitizing



Proof-of-Concept

- Part of the results being integrated in ETSI SAI WI-011 “Security aspects of using AI/ML techniques in telecom sector”

- One selected scenario worked-out as proof-of-concept
  - Day 2: Operations and maintenance
  - Task(s):
    - Reasoning, events analysis,
    - Course-of-Action execution
  - AI: LLM + tooling
    - error msg in, explanation out
    - agency: explanation is actionable
  - AdvAI: Prompt manipulation/AI-supply chain attack
    - **Not** detectable by current malware analysers etc.

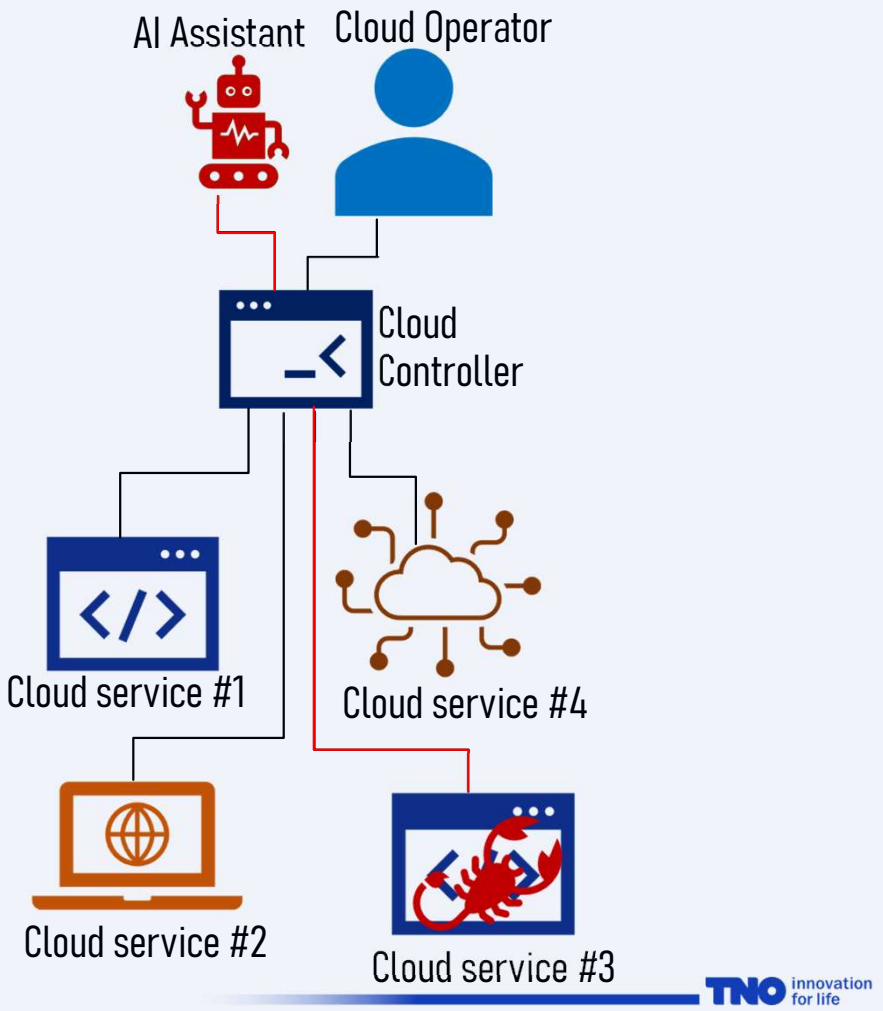


NATO OTAN
NATO Communications and Information Agency
NCI AGENCY

TNO innovation for life

Proof-of-Concept scenario

- Human Cloud Operator deploys various services in edge/tactical cloud
  - Kubernetes as cloud operating system
- AI Assistant provides troubleshooting capabilities
  - Analysis and explanation (Llama + k8sgpt)
  - Taking action, based on above analysis (TNO)
- However, AI Assistant is also new attack surface
  - Adversary poisons software update
  - Malicious instructions reach AI Assistant
  - Classis antivirus cannot detect this threat



NATO OTAN
NATO Communications and Information Agency
NCI AGENCY

TNO innovation for life

Next steps: make AI in ICT more secure

- Plans for 2025:

  1. Detecting attacks against AI in ICT context
  2. (Semi-autonomous) mitigation and response to detected attacks

- Contact
  - piotr.zuraniewski@tno.nl
  - konrad.wrona@ncia.nato.int