Adversarial AI in ICT infrastructures

C5ISR workshop (2025-03-25)

Piotr Zuraniewski, TNO

Let me introduce myself

- Senior Scientist at Networks Dept., previously Cyber Security Dept. at TNO
- Frequently project technical leader
- PhD in applied mathematics
- Former Cisco Academy Instructor and Cisco Certified Network Professional

- Interest in:
  - Programmable infrastructures
  - AI-based orchestration & management
  - Adversarial AI & AI security
  - Autonomous security response
  - Standardization (ETSI *Securing AI* delegate)
  - Supervising students & onboarding new colleagues

If and where AI is used

TNO

Not TNO

- Do you/your team use AI in your professional activities?
- If so:
  - What type of AI do you use (e.g., predictive AI, image recognition, Large Language Models,…)
  - What is the maturity level of AI usage (scale 1-5), e.g.:
    - lab/experimentation – 1
    - training/exercises – 3
    - every day's operations – 5
  - For which tasks do you use AI (scale 1-5), e.g.:
    - side/minor tasks – 1
    - regular tasks – 3
    - core tasks – 5

Why AI is not being used

TNO

Not TNO

- Do you/your team use AI in your professional activities?
- If not but you would like to:
  - What prevents you from doing so ? (please name 1 – 3 biggest obstacles)

- If not and you would not like to:
  - Please explain briefly why

AI security

- Do you recognize AI-specific risks related to your organization?
  - If so, name top 3

- If you use AI, do you manage AI related risks ?
  - If so, do you use specific framework or standard?

TNO

Not TNO

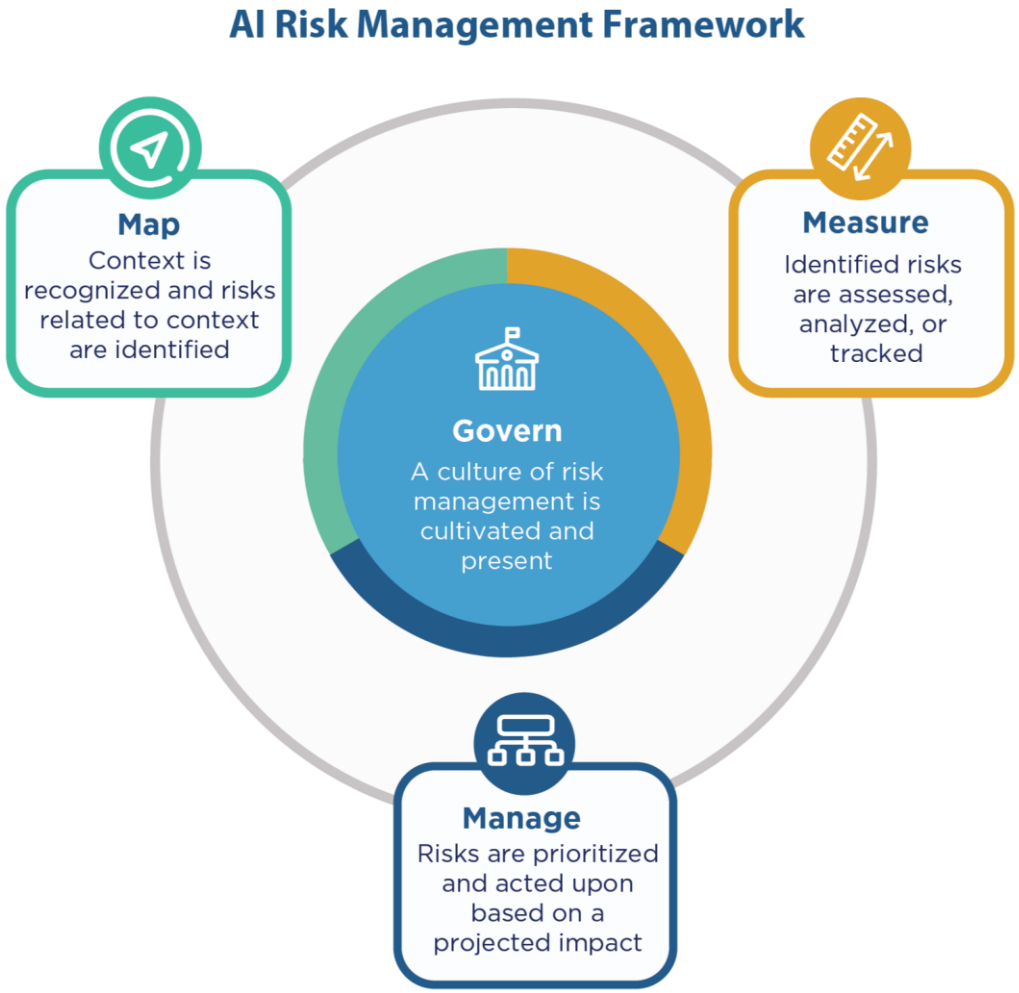


Figure: NIST AI RMF Playbook
https://airc.nist.gov/airmf-resources/playbook/

Programmable ICT infrastructures

- ICT infrastructures as a foundation for digital transformation
  - Private/public/hybrid clouds
  - Mobile networks (5G)
  - Protected Core Network (concept in Federated Mission Networking)

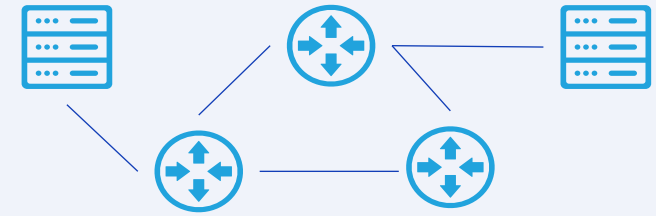
- Modern infrastructures are programmable and model-based…
- … which allows for automation of deployment, management and orchestration

GitLab

```
1  - name: pcs-tno-02-fw-tac1
2    flavour: "2cores_2GB-RAM_40GB-HDD"
3    ports:
4      - network: tac1-towards-cc-outer
5        address: 172.101.0.2
6      - network: smc-lan
7        address: 172.110.0.45
```

Descriptor of 'skeleton PCN infrastructure' (snippet)

Pulumi + openstack CLOUD SOFTWARE

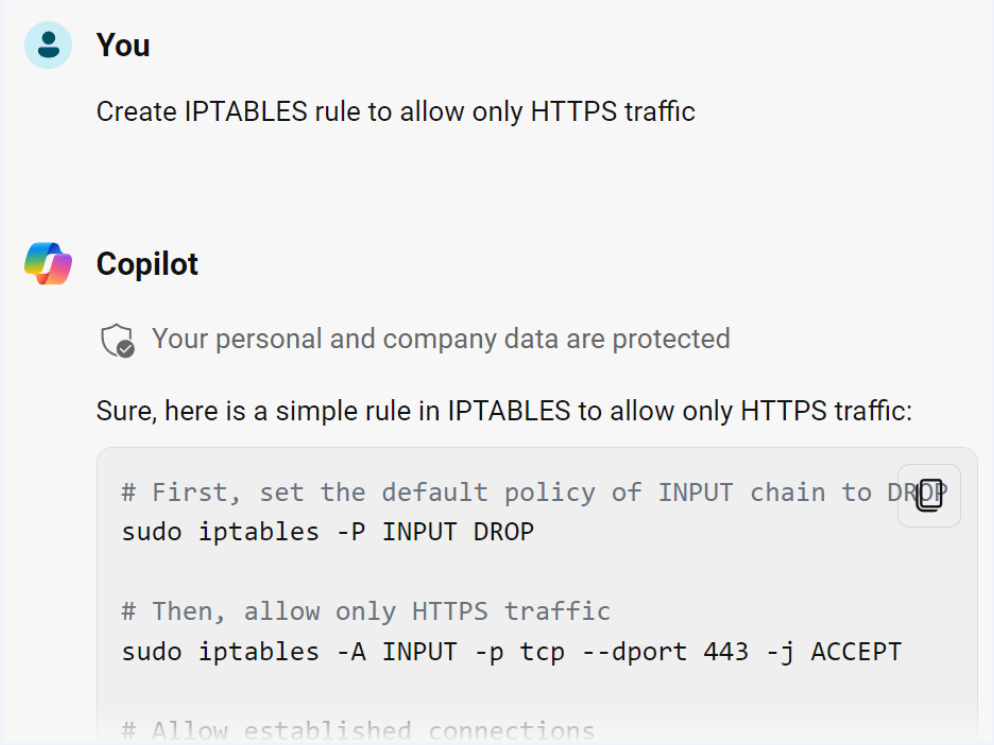


Instantiation of PCN 'skeleton infrastructure' in the cloud

AI everywhere – also in ICT infrastructures

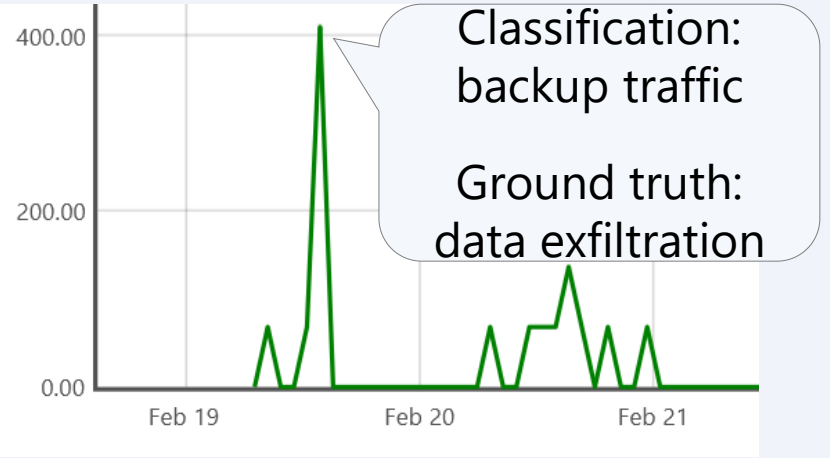
- Vision: next-generation infrastructures will be:
  - intent-based,
  - self-managing,
  - self-organizing
  - self-defending.

- Behind each ‘self-*’ feature, there will be AI system
  - Predictive AI: anomaly detection, event classification,…
  - Generative AI: log analysis, playbook generation,…

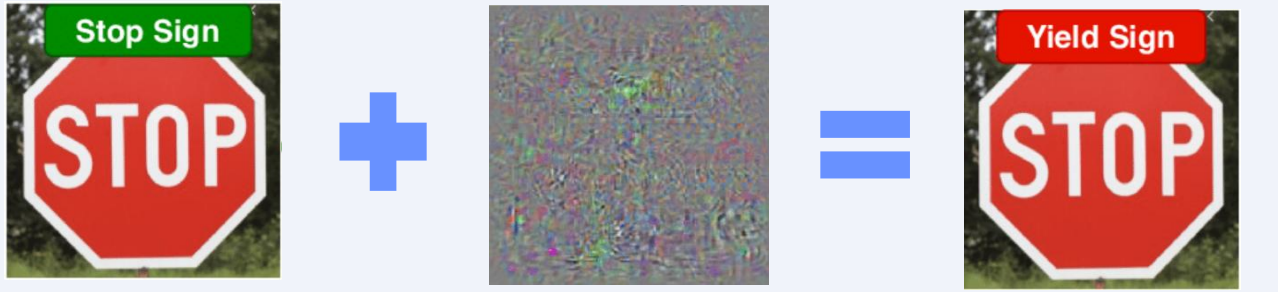


AI poses new type of security problem in ICT infra

- **AI systems can become targets on their own**
- New, “data-centric” type of attack, different from current cybersecurity problems
  - Poisoning training data may lead to attack classified as good traffic
  - Prompt injection may lead to malicious code execution



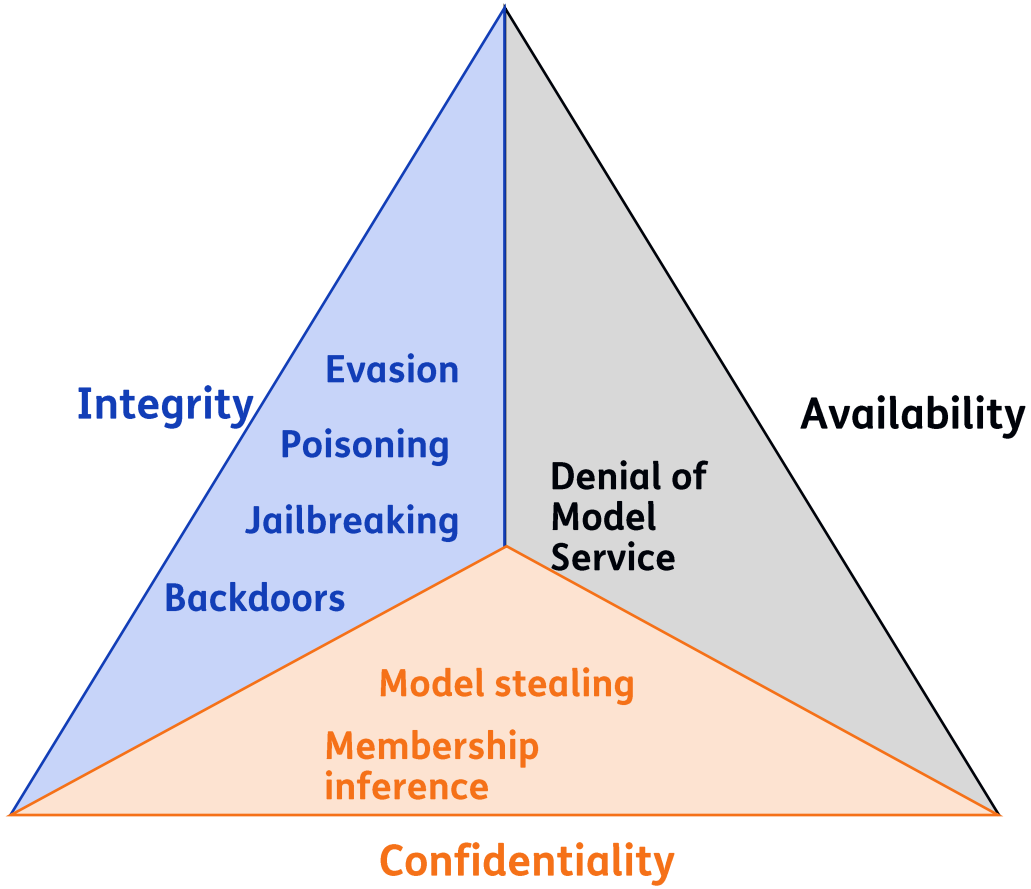
- **No systematic solution as of now**
  - Most research in image recognition (stop sign example)
  - Less so in ICT infra use cases



Authentic Input | Adversarial Perturbation | Adversarial Input

Stop sign example inspired by: https://ai-agora.uib.no/seminar/martha-larson-ai-vs-our-eye-new-perspectives-on-adversarial-images-10-march-13-15-14-30/

The CIA triad for AI model and application security



TNO activities in AI security within ICT infra context

- ADVICE – Adversarial AI in ICT Infrastructures
  - 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 (LLM) 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

6G FUTURE NETWORK SERVICES

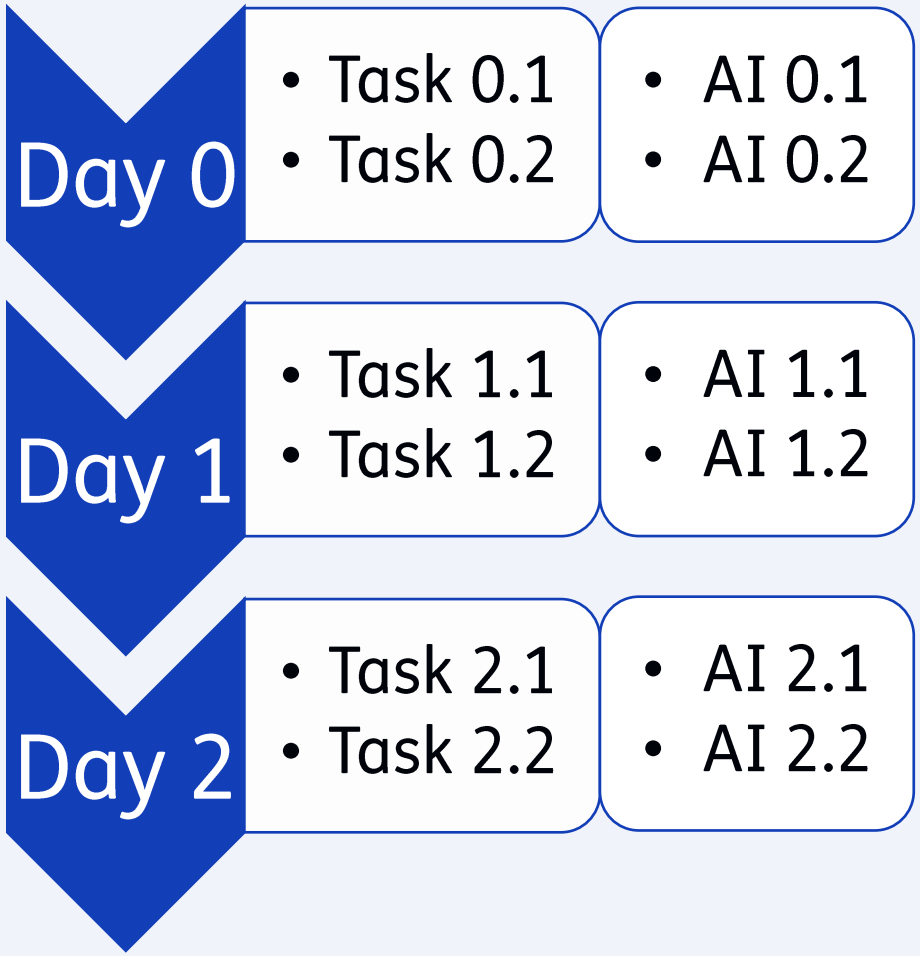
Rijksoverheid

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

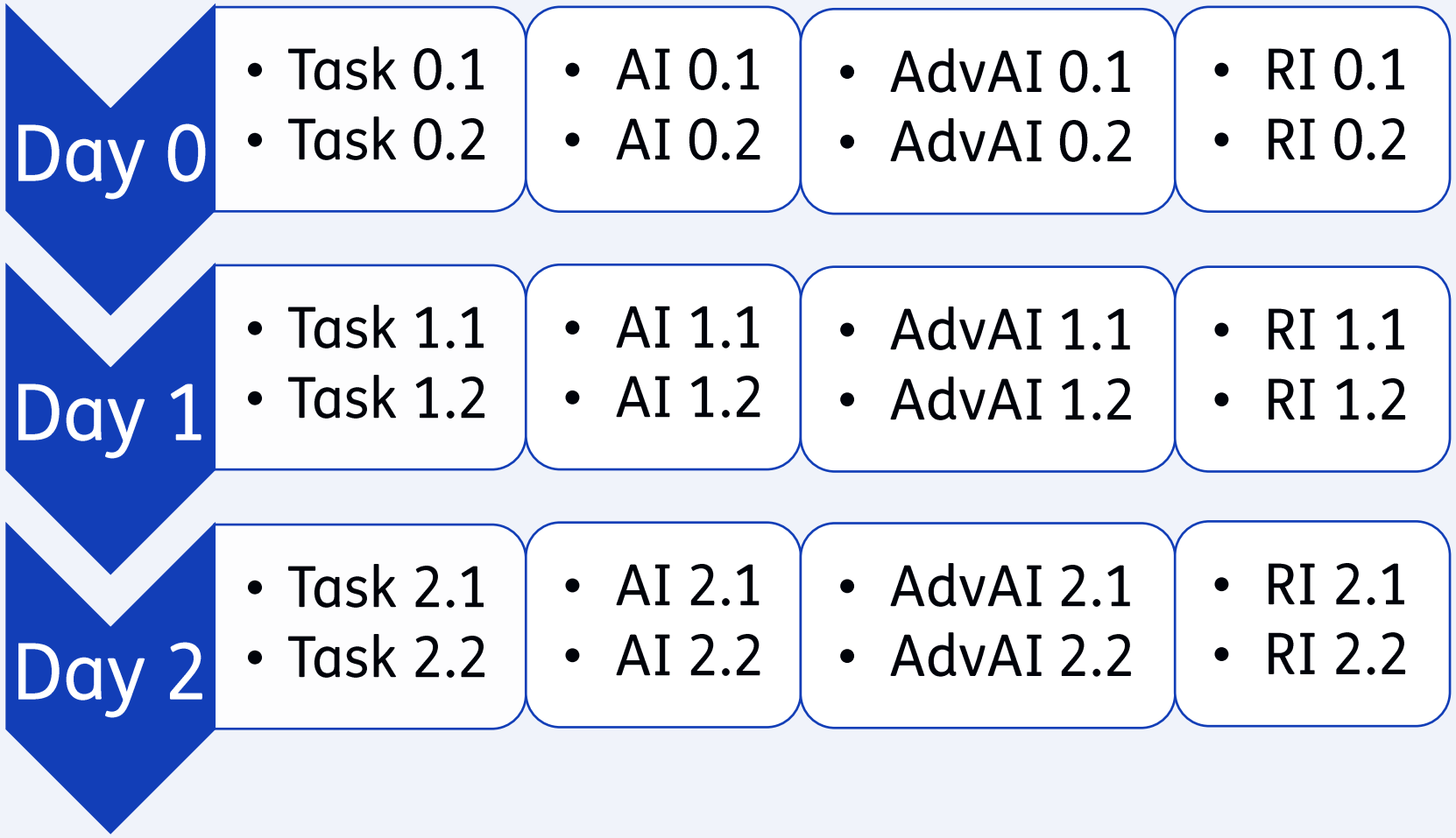


[ETSI_NFV022] ETSI GR NFV-EVE 022 „Network Functions Virtualisation (NFV) Release 5; Architectural Framework; Report on VNF configuration”

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

Adversarial scenarios in ICT infrastructure lifecycle

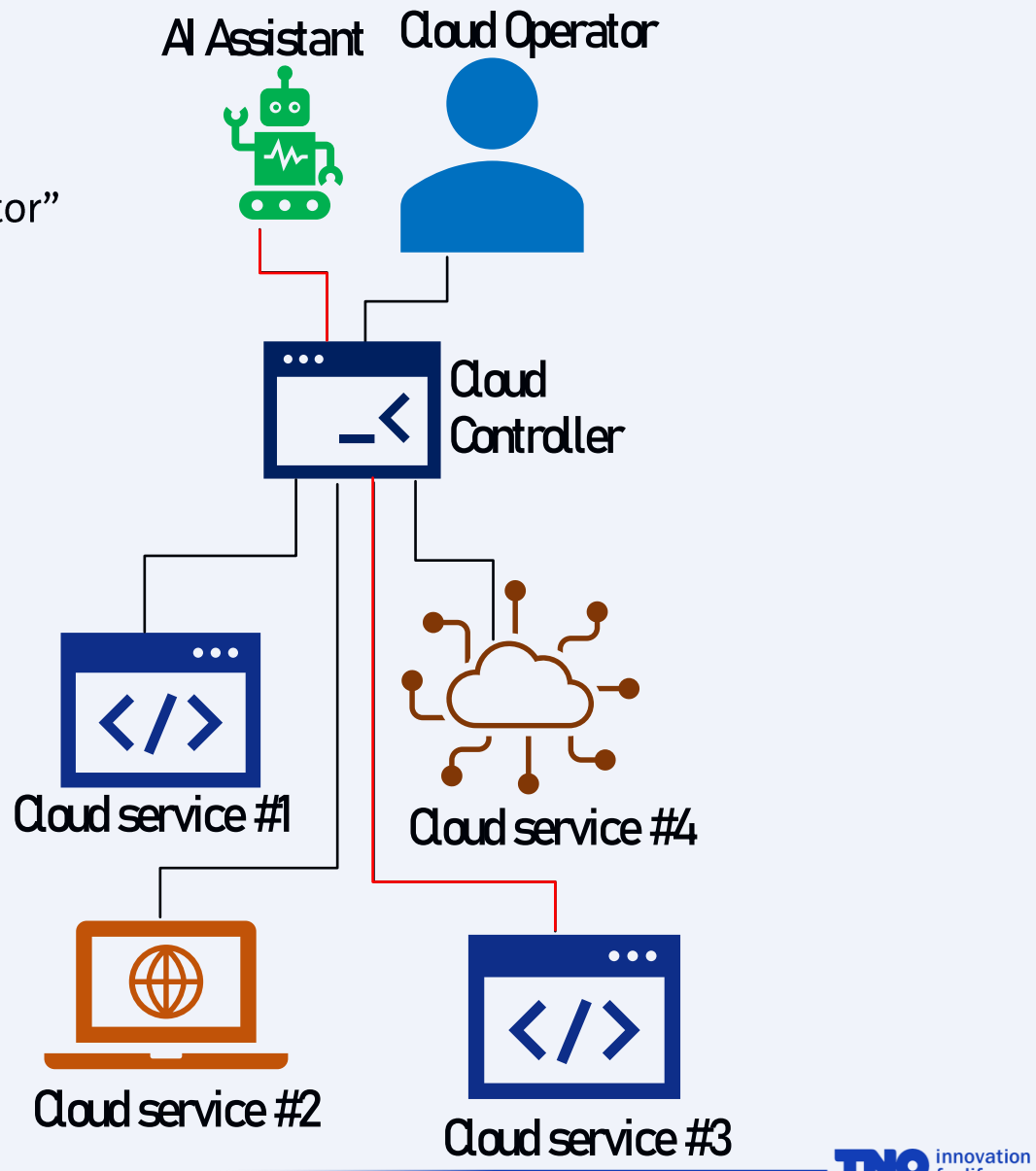
- Next, for each AI technique, identify **adversarial** AI technique
  - Example (cont'd):
    - AI: use LLM
    - **AdvAI: Data extraction**
- Attempt to assess 'risk index' (RI) using e.g., CVSS4
- Consider mitigation measures, both:
  - 'classic' e.g., access control
  - AI-specific e.g., prompt sanitizing



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.

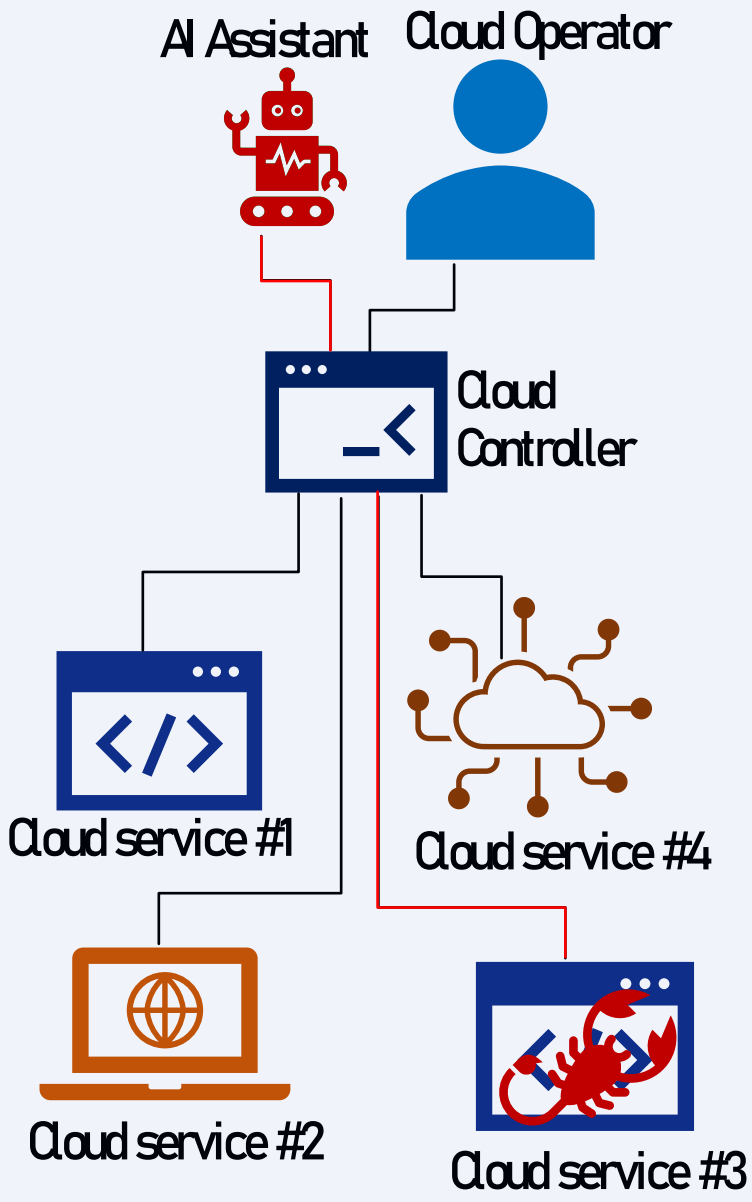


Proof-of-Concept scenario

- 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)

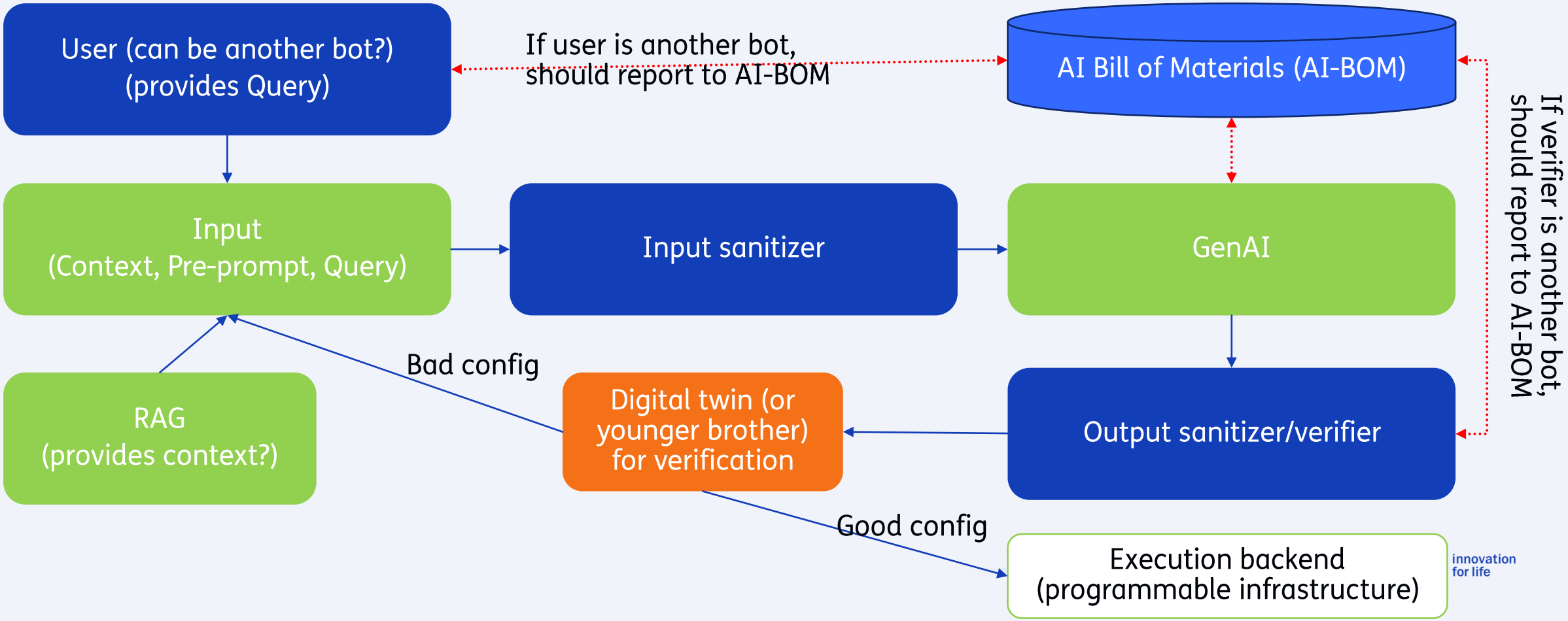
- However, AI Assistant is also new attack surface
  - Adversary poisons software update
  - Malicious instructions reach AI Assistant
  - Classis antivirus cannot detect this threat



FNS

Better workflow ? (W.I.P.)

- Both unintentional and intentional harm possible – sanitize both input and output for/of GenAI
- Verify/validate before deploying in production, if problems/errors, cycle back to LLM for corrected config
- Be mindful of what AI you employ (AI-BOM)



AI Bill of Materials (AI-BOM) … but S-BOM first

- Point of departure: Software Bill of Materials (S-BOM)
  - Inventorize software (versions, licenses, libraries, dependencies,…)
  - Store information in machine readable format

- S-BOM also helps in mitigating security risks:
  - List of know vulnerabilities, also from 3rd party software
  - If new vulnerability is disclosed, SoC knows if/where the problem is

- Two well known examples:
  - SPDX (Linux Foundation)
  - CycloneDX (OWASP)[1]


```
{
    "bom-ref": "brick/math-0.9.3.0",
    "type": "library",
    "name": "math",
    "version": "0.9.3",
    "group": "brick",
    "description": "Arbitrary-precision arithmetic library",
    "licenses": [
        {
            "license": {
                "id": "MIT"
            }
        }
    ],
    "purl": "pkg:composer/brick/math@0.9.3",
    "externalReferences": [
        {
            "type": "distribution",
            "url": "https://api.github.com/repos/brick/math/zipball/ca57
        },
        {
            "type": "vcs",
            "url": "https://github.com/brick/math.git"
```


[1] SBOM snippet from bom-examples/SBOM/laravel-7.12.0/bom.1.4.json at master · CycloneDX/bom-examples · GitHub

AI Bill of Materials (AI-BOM)

- Builds on top of S-BOM concept: inventorize components of AI system
- Many fields in AI-BOM can be identical to S-BOM (package version, location,...)
- However, AI-specific information can also be captured
  - Type of model, data preprocessing steps, human-in-the-loop,..
  - Information about data used to train the model can also be part of AI-BOM
    - If data source was poisoned, AI-SoC knows if/where the problem is

- Both SPDX and CycloneDX have AI/ML modules
- AI-BOM can help in becoming compliant with regulations
  - E.g., SPDX offers mapping of AI Act clauses to AI-BOM features, see link

Sub-categories	EU AI Act description and clause	In AI BOM?	Matching field in AI BOM	SPDX profile
System classi-fication	"A short summary of the grounds on which the AI system is considered to be not-high-risk in application of the procedure under Article 6(3)" - Annex VIII Section B (7)	✔	informationAboutApplication; hasDocumentation relationship	AI, Core

Note: while SPDX v2.2.1 became ISO standard, AI BOM was added only in v.3.0.0, https://www.iso.org/standard/81870.html

Note: CycloneDX v.1.6 (current stable, supporting ML BOM) was published as ECMA standard https://ecma-international.org/publications-and-standards/standards/ecma-424/

Next steps: make AI in ICT infra more secure

- Plans for 2025:

  1. Detecting attacks against AI in (military) ICT infrastructures context (->IST-221)
  2. (Semi-autonomous) mitigation and response to detected attacks

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