STARLIGHT

Sustainable Autonomy and Resilience for
LEAs using AI against High Priority Threats

AI-based cybersecurity and protection of LEA AI solutions

Campus Cyber, 2024-06-10

Cédric Gouy-Pailler & Nizar Touleimat - CEA

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101021797

General objectives

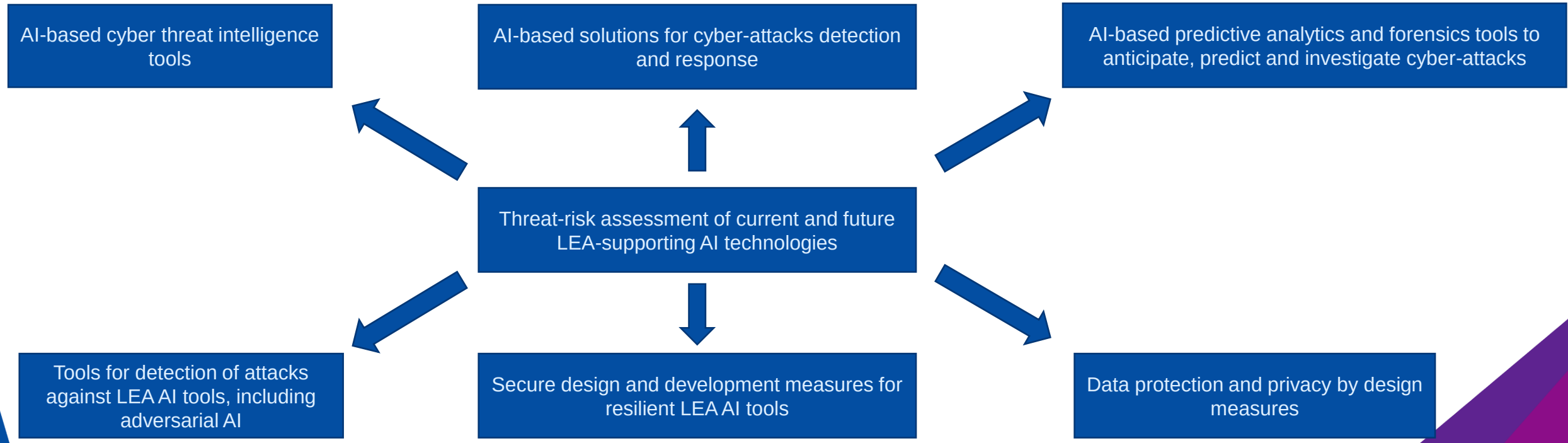
- **AI for cybersecurity:** Providing LEAs with AI-based tools to
  - analyze,
  - predict,
  - anticipate
  - mitigate

  } **present** and **future** cyber-attacks

- **Cybersecurity of AI**

  Building { methodologies / technologies } to ensure **robustness**, **security** and **resilience** of **AI tools** already in use, or to be used by LEAs

Topics of interest



From research institutions to agencies

Research

LEA / technical partners co-developments

Availability for LEA

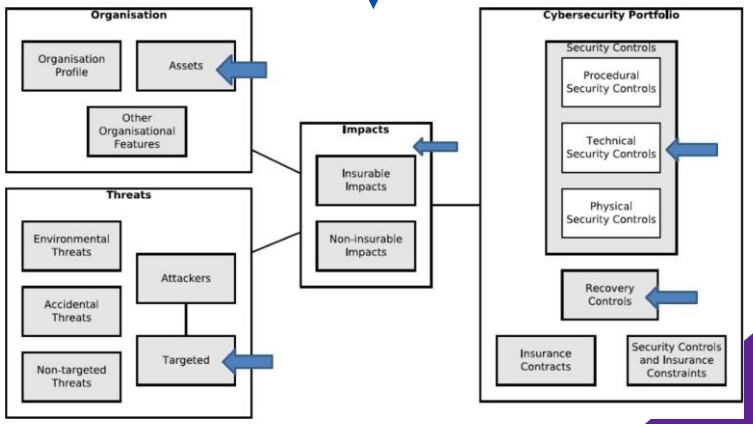
Threat-risk assessment of current and future LEA-supporting AI technologies

qAIScore
Qualitative multicriteria analysis based on ALTAI

AIRMF
Quantitative analysis of risks and impacts

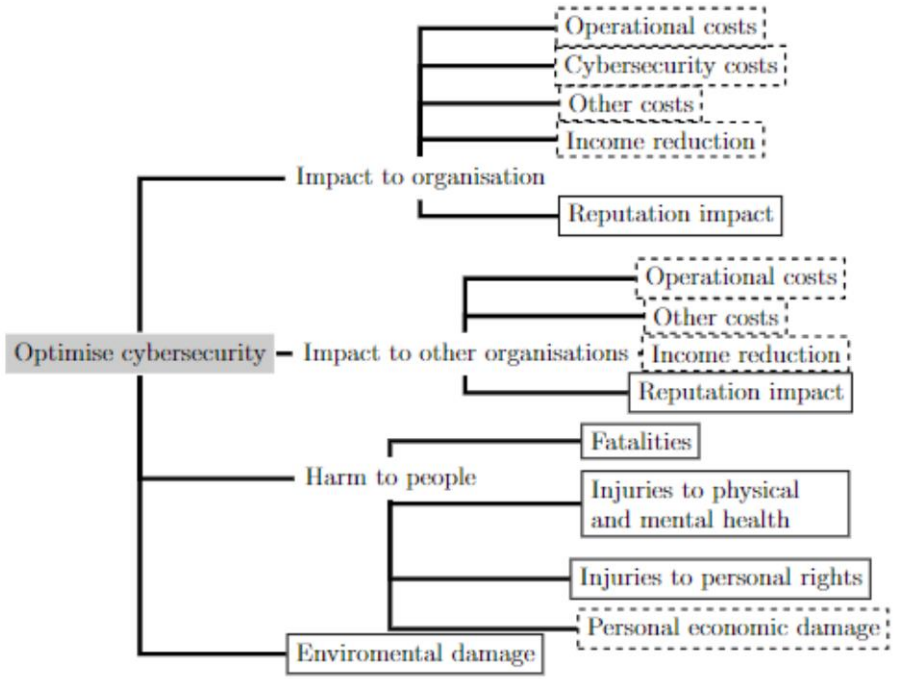
The evaluation and the comparison follow a qualitative scoring that can be used to:

- Reject AI tools that receive low scores. A qualitative analysis may already provide enough information to flag a tool as too risky.
- Understand that deeper analysis is required to understand risks and impacts when a qualitative analysis is too limited.
- Accept AI tools.



Threat-risk assessment of current and future LEA-supporting AI technologies

- J. M. Camacho, A. Couce-Vieira, D. Arroyo, et D. R. Insua, « A Cybersecurity Risk Analysis Framework for Systems with Artificial Intelligence Components ». arXiv, 3 janvier 2024. doi: 10.48550/arXiv.2401.01630.



A Cybersecurity Risk Analysis Framework for Systems with Artificial Intelligence Components

J.M. Camacho[1], A. Couce-Vieira[2], D. Arroyo[3], and D. Rios Insua[1]

[1]Inst. Mathematical Sciences, ICMAT-CSIC, Spain

[2]University of Vigo, Spain

[3]Inst. Information Technology, ITEFI-CSIC, Spain

Focus on « A Cybersecurity Risk Analysis Framework for Systems with Artificial Intelligence Components »

Novel Risk Analysis Issues

- Emergent AI capabilities and risks discovered post-deployment.
- Increased risks from AI's complexity.
- New challenges in explainability, reliability, and safety.

AI-related Impacts and Assets

- Broader vision of cybersecurity impacts beyond technical aspects.
- Inclusion of AI-specific impacts: accuracy, reliability, safety, and privacy.
- Adaptation of traditional cybersecurity frameworks to AI impacts.

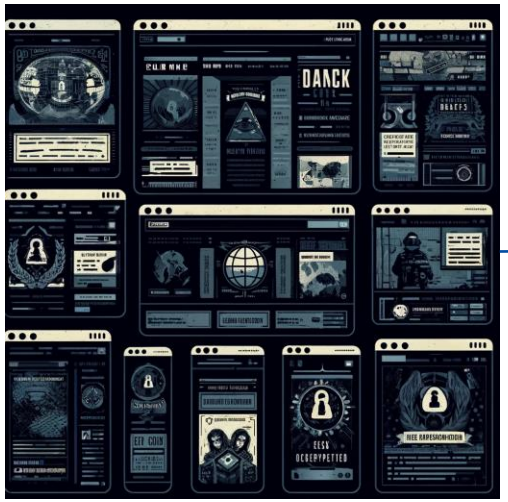
Algorithm 3 Attack simulation in a general facility

Input: G, Λ, distributions for P,Q,Q^s

for $j \in \{1, \ldots, M\}$ **do**
  Generate number N of attacks using Λ
  for $i \in \{1, \ldots, N\}$ **do**
    Generate P, Q, Q^s
    Simulate attack transit using P,Q, Q^s ▷ Algorithm 1
    Simulate impact l_i from attack ▷ Algorithm 2
  Aggregate $l^j = \sum_{i=1}^{N} l_i$
Return: $\{l^j\}_{j=1}^{M}$

AI-based cyber threat intelligence tools

Dark Web Monitor



Label

- Dark Web Monitor
  - Content classification module developed

Context-Aware Feature Vectors in Dark Web Page Classification

Sander Brinkhuijsen[1,2],
Supervisors: Romana Pernisch[1,3] Eljo Haspels[2], and Mark van Staalduinen[2]

[1] Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
[2] CFLW Cyber Strategies, The Hague, the Netherlands
[3] Discovery Lab, Elsevier, Amsterdam, the Netherlands
sander.brinkhuijsen@cflw.com, r.pernisch@vu.nl, eljo.haspels@cflw.com, mark.vanstaalduinen@cflw.com

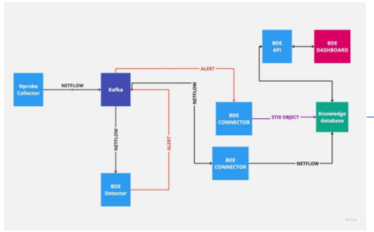
Codevelopment example 1 – lead by CEA & MinInt

- Log analysis to identify hacker source and compromised accounts

Technical Lead:	CEA
LEA lead:	MinInt
Interested/involved technical partners:	Pluribus One VicomTech ITTI
Interested/involved LEA partners:	Belgium Police
Others:	CNRS-IRIT, DCIS

Tools

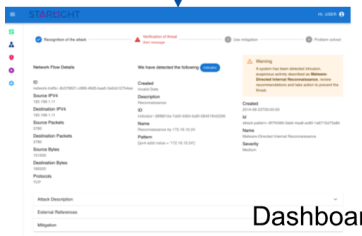
TCAD (Trustworthy Cyber-Attack Detector)
Netflow data analysis for detection of hidden patterns of malicious activities



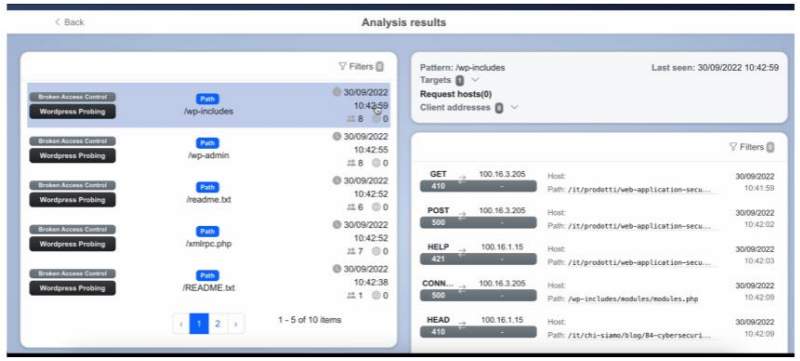
Architecture for detection

Precision results

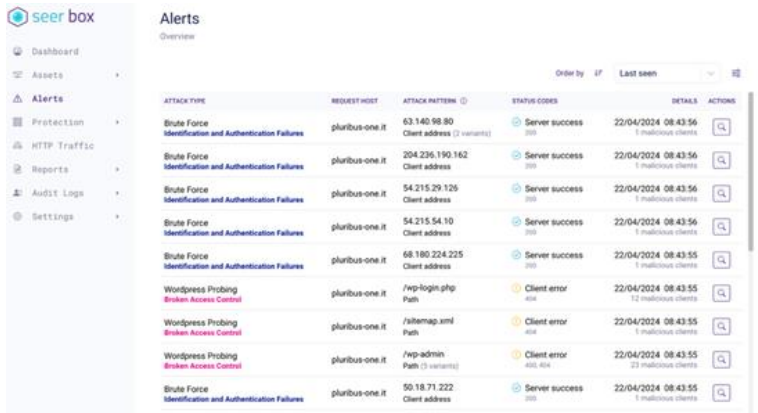
Random Forest – Spark TCAD				
	ACC	Precision	Recall	F1
Port Scanning	1	0.99	0.99	0.99
DoS	1	0.99	0.99	0.99
Brute Force	0.95	0.96	0.98	0.96

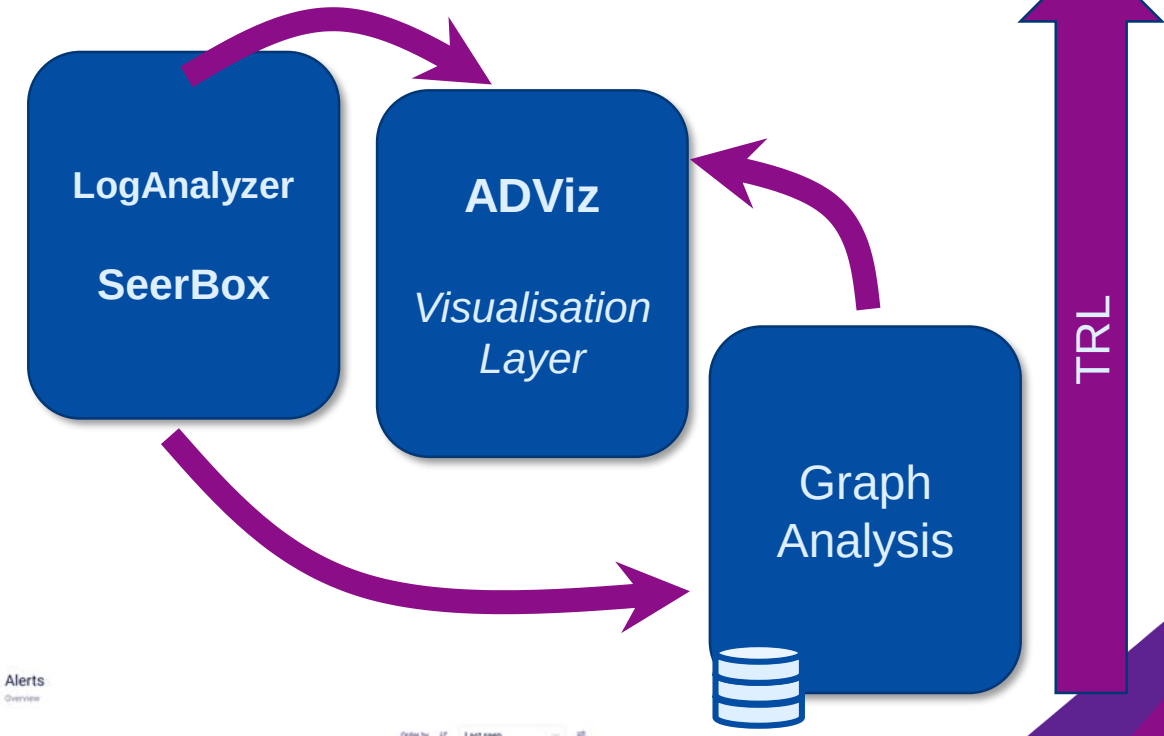


Dashboard

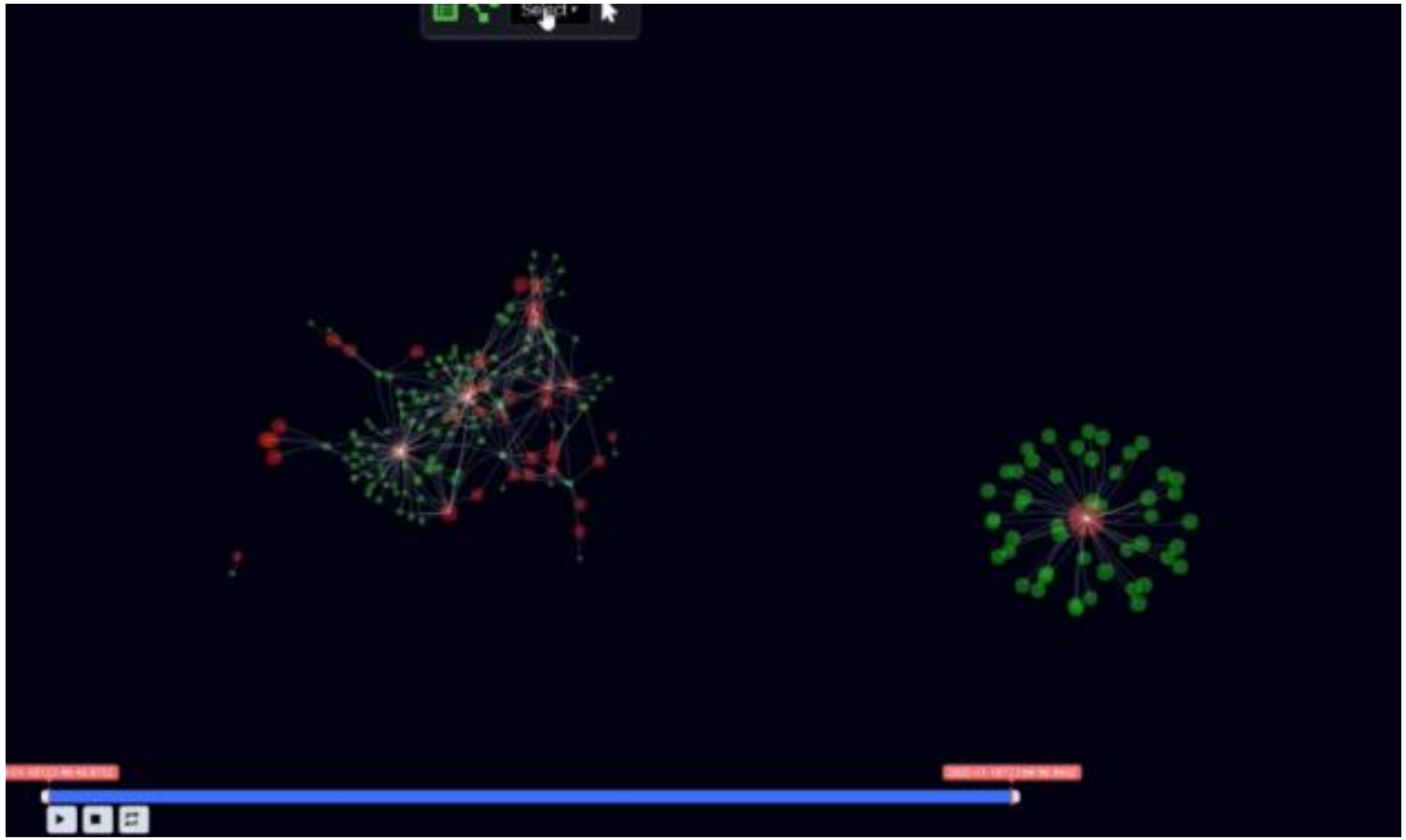


Log Analyzer
Pluribus tool





Adviz tool developed at CEA (live démo)



Codevelopment example 2 – lead by INOV & PJ

- Forensics: identification of cyberattacks in post-mortem systems

Technical Lead:	INOV
LEA lead:	PJ
Interested/involved technical partners:	
Interested/involved LEA partners:	
Others:	

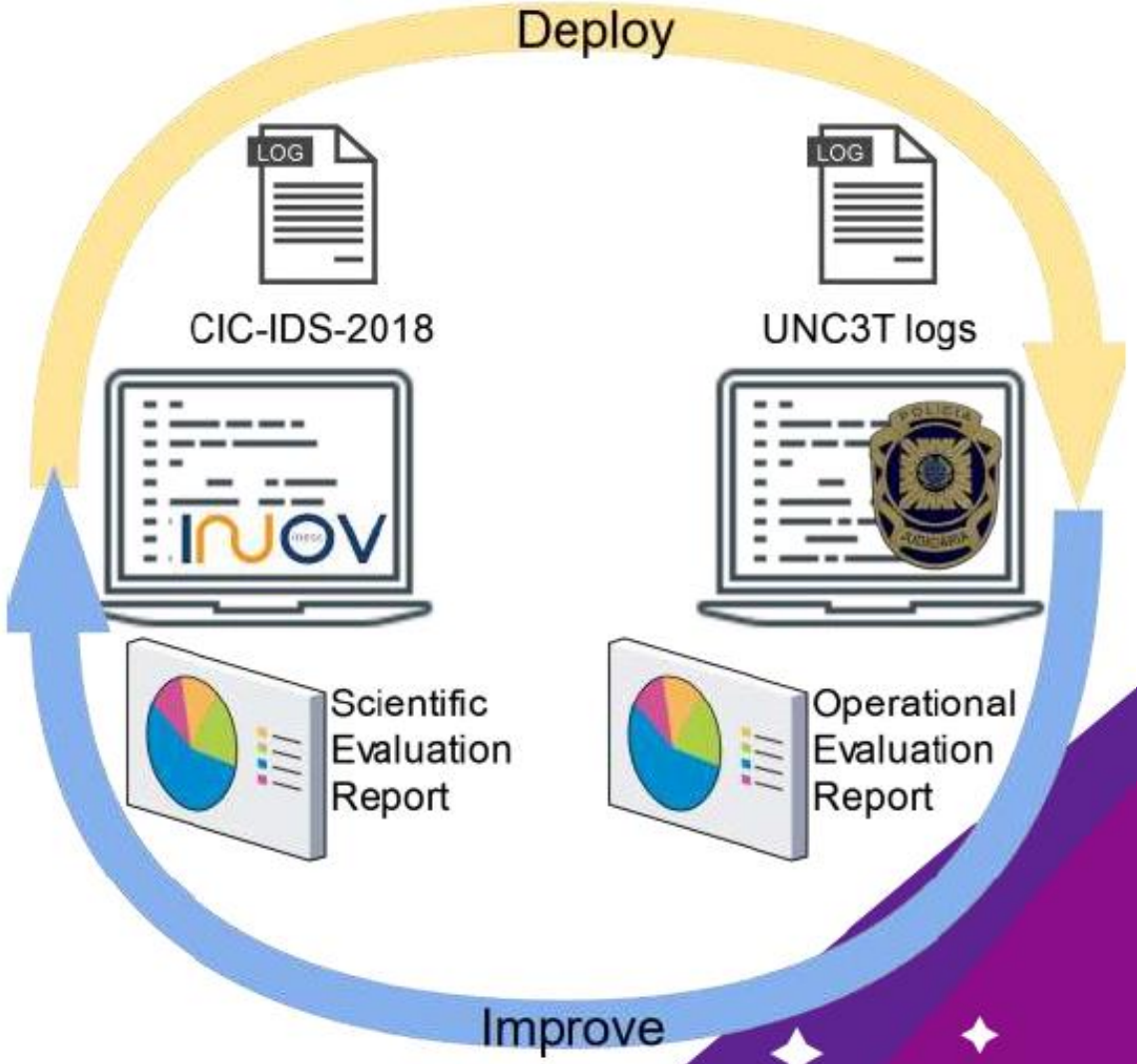
Detection of Cyberattacks in Post-mortem systems

- In cases of post-mortem systems, **LEA investigators and forensic experts need** to **analyse logs** and interview security administrators, to **discover evidences** to prosecute the attacker.
- **Requires knowledge** of the affected system
- Insights gathered from **interviews with the affected ICT organization**
- Reconstitute the **timeline of the attack** and **modus-operandi**

Methodology

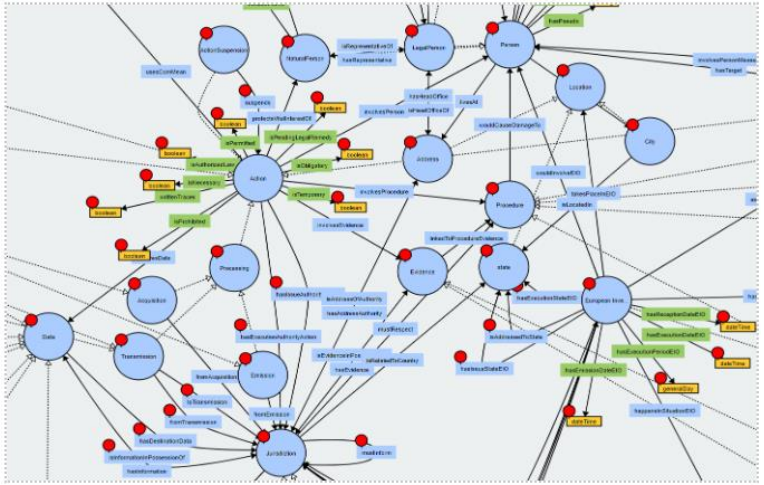
Roadmap

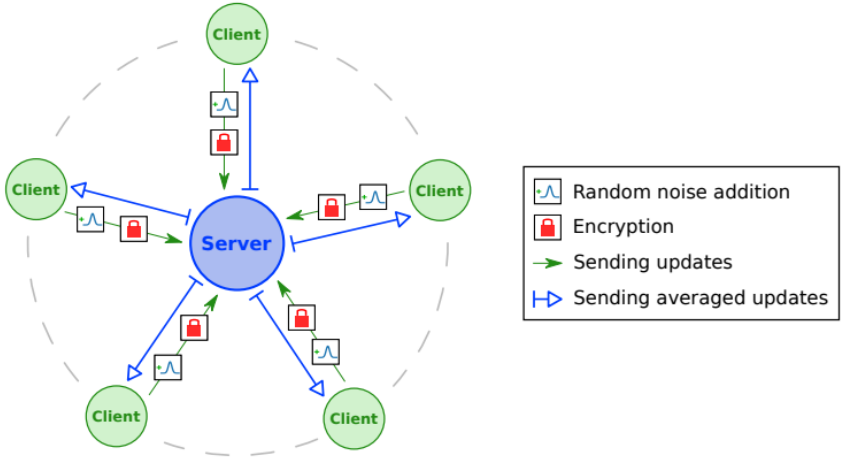
- INOV develops synthetic data from widely used CIC-IDS-2018
- PJ validates with UNC3T open cases
- 8 Sprint releases from january to september (next GA)



Data protection and privacy by design measures

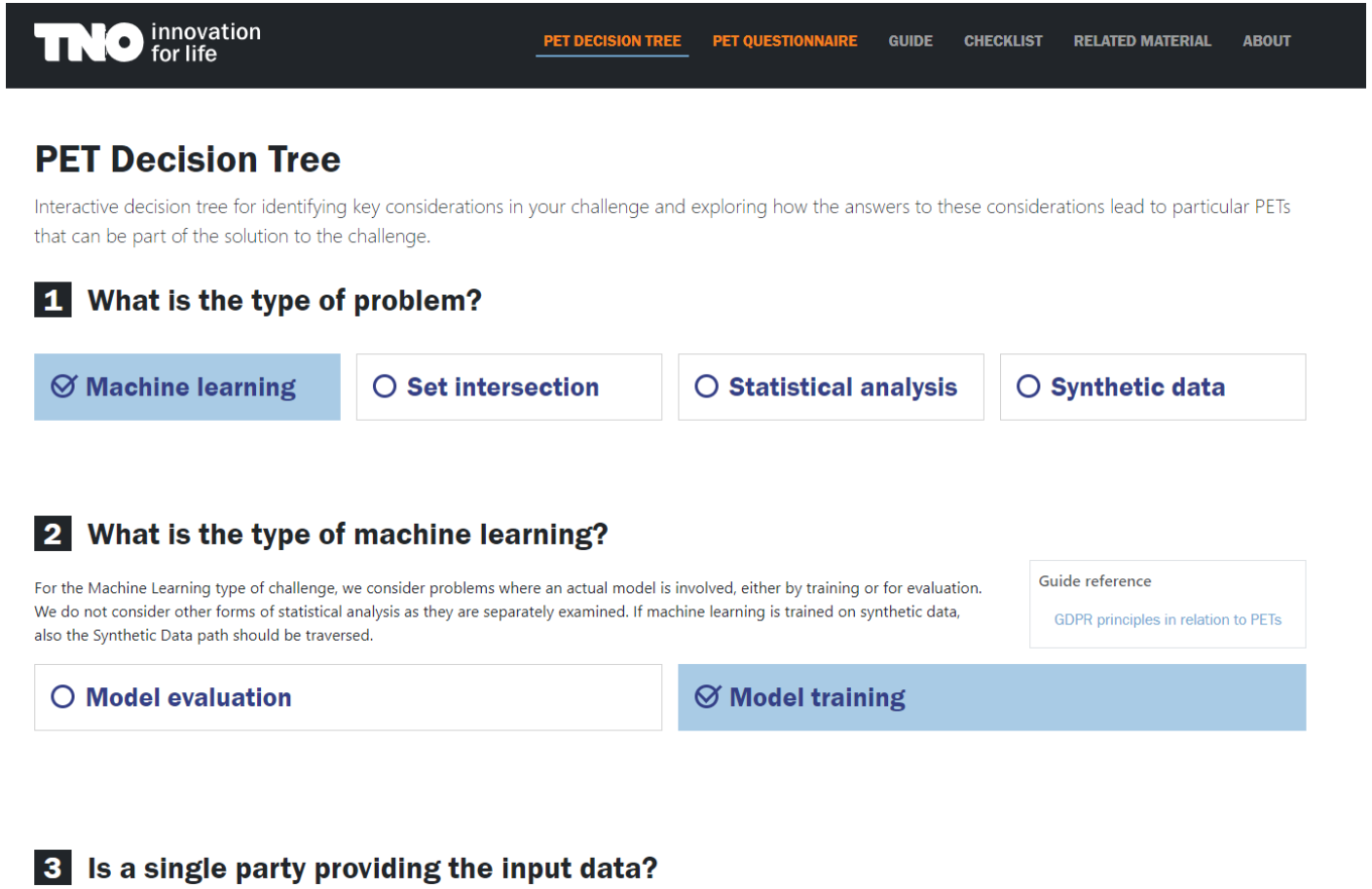
- Tool 1: design of a tool to support knowledge sharing among LEAs
  - Progress towards transforming regulations into formal rules
- Tool 2: design of a private-by-design AI training among LEAs
  - Federated learning for training without sharing
- Tool 3: decision tool to choose the adapted privacy-enhanced technology from objectives and constraints





Privacy-enhancing technologies

https://tno-pet-explorer.online/decision-tree/



Public material

- G. Saraiva, F. Apolinário, et M. L. Pardal, « IM-DISCO: Invariant Mining for Detecting IntrusionS in Critical Operations », in *Computer Security. ESORICS 2023 International Workshops*, S. Katsikas, H. Abie, S. Ranise, L. Verderame, E. Cambiaso, R. Ugarelli, I. Praça, W. Li, W. Meng, S. Furnell, B. Katt, S. Pirbhulal, A. Shukla, M. Ianni, M. Dalla Preda, K.-K. R. Choo, M. Pupo Correia, A. Abhishta, G. Sileno, M. Alishahi, H. Kalutarage, et N. Yanai, Éd., Cham: Springer Nature Switzerland, 2024, p. 42-58. doi: 10.1007/978-3-031-54129-2_3.
- M. Pawlicki, A. Pawlicka, R. Kozik, et M. Choraś, « Advanced insights through systematic analysis: Mapping future research directions and opportunities for xAI in deep learning and artificial intelligence used in cybersecurity », *Neurocomputing*, vol. 590, p. 127759, juill. 2024, doi: 10.1016/j.neucom.2024.127759.
- J. M. Camacho, A. Couce-Vieira, D. Arroyo, et D. R. Insua, « A Cybersecurity Risk Analysis Framework for Systems with Artificial Intelligence Components ». arXiv, 3 janvier 2024. doi: 10.48550/arXiv.2401.01630.
- D. Rios Insua, R. Naveiro, V. Gallego, et J. Poulos, « Adversarial Machine Learning: Bayesian Perspectives », *Journal of the American Statistical Association*, vol. 118, nº 543, p. 2195-2206, juill. 2023, doi: 10.1080/01621459.2023.2183129.
- R. Naveiro, D. R. Insua, et J. M. Camacho, « Augmented probability simulation for adversarial risk analysis in general security games », in *12th International Defence and Homeland Security Simulation Workshop, DHSS 2022*, 2022. Consulté le: 9 juin 2024. [En ligne]. Disponible sur: https://digital.csic.es/handle/10261/346388
- V. Gallego, R. Naveiro, A. Redondo, D. R. Insua, et F. Ruggeri, « Protecting Classifiers From Attacks. A Bayesian Approach ». arXiv, 18 avril 2020. Consulté le: 9 juin 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2004.08705
- S. Brinkhuijsen, « Context-Aware Feature Vectors in Dark Web Page Classification », Consulté le: 9 juin 2024. [En ligne]. Disponible sur: https://bnaic2022.uantwerpen.be/wp-content/uploads/BNAICBeNeLearn_2022_submission_8031.pdf

STARLIGHT

Sustainable Autonomy and Resilience for LEAs using AI against High Priority Threats

Questions?

Cédric Gouy-Pailler

CEA

cedric.gouy-pailler@cea.fr

https://starlight-h2020.eu | @starlight_h2020 | starlight@cea.fr

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101021797