Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

SERICS
SECURITY AND RIGHTS IN THE CYBERSPACE

Cybersecurity threat elicitation under human-artificial intelligence sources of risk

Spoke 06

Giampaolo Bella, Mario Raciti,

Simone Di Mauro

SERICS
SECURITY AND RIGHTS IN THE CYBERSPACE

Third review meeting
Second Software and Platform Security Workshop
26-28 June 2025

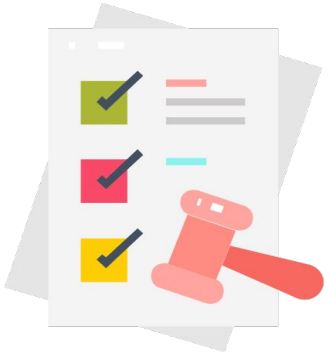
Università di Catania

IMT SCUOLA ALTI STUDI LUCCA

Introduction

Threat modelling must map system safeguards to complex, *multi-domain regulations* to ensure **legal compliance**

Manual extraction of requirements from lengthy *legislative texts* is **slow and error-prone**

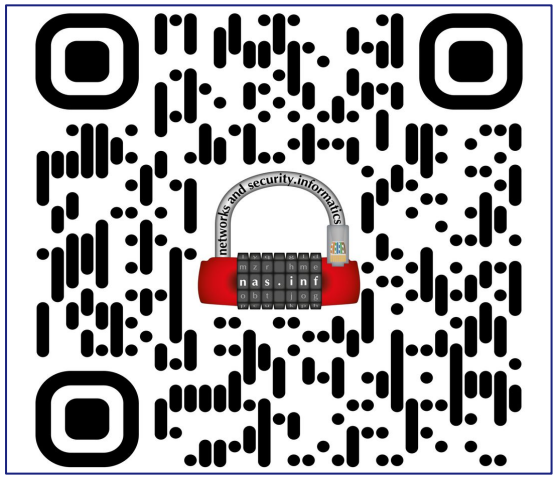


UniCT's Role (SCAI WP1: T1.2 - Innovative techniques for evaluating cyber risk)

Definition of a **cyber-risk assessment methodology** for *heterogeneous* ICT infrastructures, including **AI-** and **human-derived threats**

Goal i: Methodology definition (D4)

Goal ii: Vertical application (D5)



D4 at a Glance

Objective: Automated, *HAI-powered* threat elicitation from **regulatory texts**

Challenges:

Volume & complexity of **multi-domain** regulations (AI Act, NIS 2, ISO 9241-210:2019, etc.)

Need to cover cyber, privacy, AI, human factors

Outcome: A **three-phase** methodology

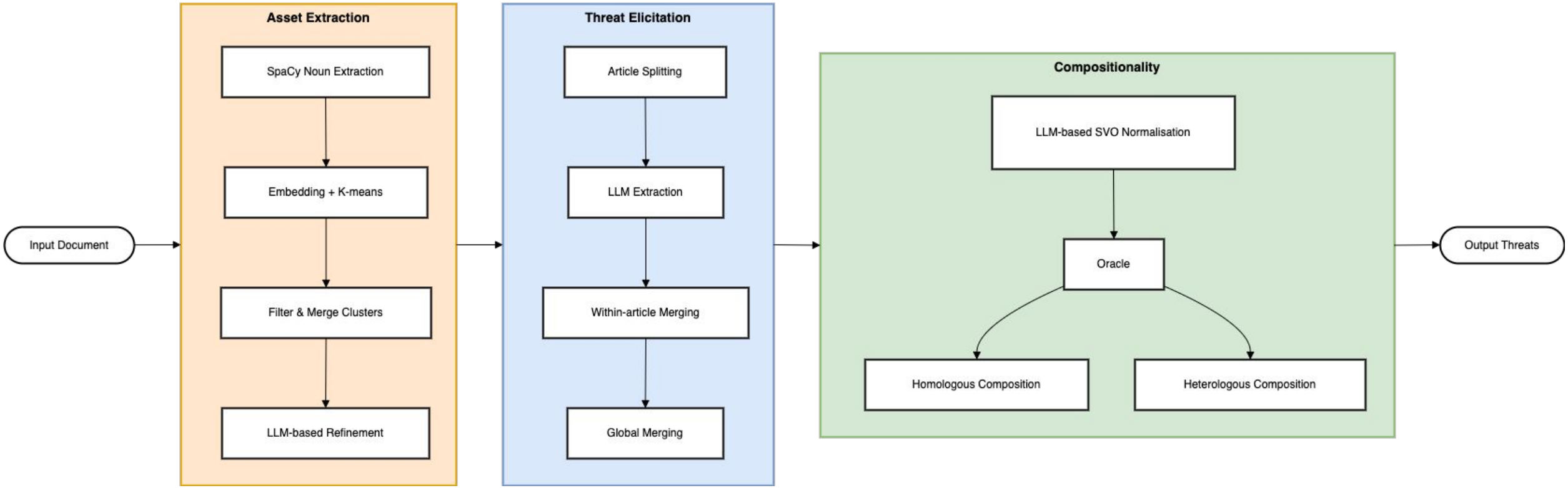
Agenda

1. Introduction
2. **Methodology**
3. Partial Validation
4. Conclusions

SERICS

SECURITY AND RIGHTS IN THE CYBERSPACE

Methodology Overview



Agenda

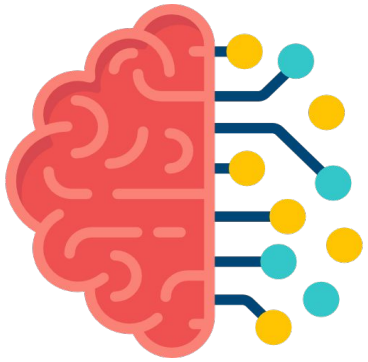
1. Introduction
2. **Methodology →**
   Asset Extraction
3. Partial Validation
4. Conclusions

SERICS
SECURITY AND RIGHTS IN THE CYBERSPACE

Asset Extraction

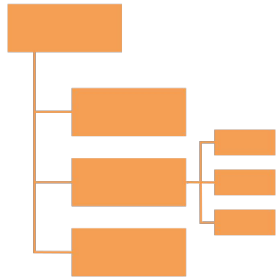
Based on *Natural Language Processing (NLP)* and *Clustering*, with an *LLM-based refinement*

SpaCy Noun Extraction → Word Embeddings → K-means → LLM Refinement



Asset Extraction - NLP

SpaCy Noun Extraction → *Word Embeddings → K-means → LLM Refinement*



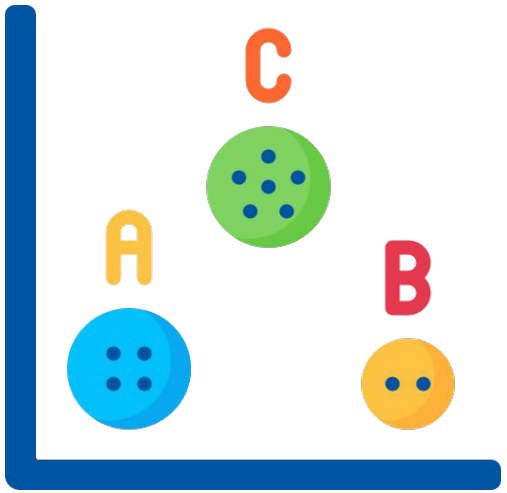
“Organizations shall ensure the integrity of personal data by implementing encryption and access controls.”

["Organizations", "integrity", "personal data", "encryption", "access controls"]

Asset Extraction - Clustering

SpaCy Noun Extraction → ***Word Embeddings → K-means*** → *LLM Refinement*

- Compute **embeddings** for each noun and run **K-means**
- Human analyst *chooses optimal k* via silhouette score → here k = 3



Cluster A: ["encryption", "access controls"]

Cluster B: ["personal data", "integrity"]

Cluster C: ["Organizations"]

Asset Extraction - Refinement

SpaCy Noun Extraction → Word Embeddings → K-means → ***LLM Refinement***

- **Set thresholds:** *st1 = 0.6* and *st2 = 0.8* for semantic similarity
- **Filter:** drop any noun whose *average similarity* to its cluster-mates < st1
- **Merge:** if two clusters' centroids cosine-sim > st2
- **Select:** LLM selects the assets (Prompt 1)

Assets cluster 1: ["access controls", "encryption"]

Assets cluster 2: ["personal data", "data integrity"] → selected as "assets"

Agenda

1. Introduction
2. **Methodology → Threat Elicitation**
3. Partial Validation
4. Conclusions

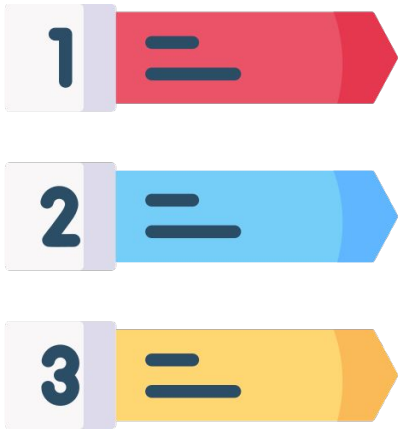
SERICS

SECURITY AND RIGHTS IN THE CYBERSPACE

Threat Elicitation

Article-Level Analysis:

- Split document into *articles*
- Human analyst choses *N*
- For each article, run LLM *N*× to extract asset–threat pairs (Prompt 2)



Consolidation:

- *Within-article* merging (Prompt 3) → reduce redundancy
- *Global* merging (Prompt 4) → unified threat list

Threat Elicitation - Before Merge

Article #	Extracted Threats
1	Unauthorised access due to missing authentication; Sensitive data exposed via public endpoints; Weak session management allowing token reuse.
2	SQL injection risk in user profile update; Lack of input validation on form fields.
3	Error messages disclose stack traces; Verbose logs reveal internal paths.

Threat Elicitation - After Merge

Article #	Consolidated Threat
1	Inadequate authentication and session controls lead to unauthorised access and potential data exposure.
2	Improper input handling exposes the system to injection attacks and unexpected behaviors.
3	Excessive error information leakage may aid attackers in understanding system internals.

Agenda

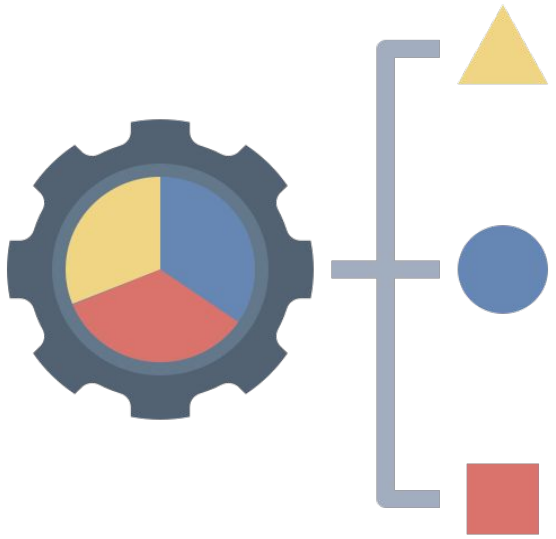
1. Introduction
2. **Methodology →** **Compositionality**
3. Partial Validation
4. Conclusions

SERICS

SECURITY AND RIGHTS IN THE CYBERSPACE

Compositionality

Goal: *Identify higher-order, cross-domain threat chains*



Normalise each threat to **SVO** form → *Oracle(s)*

Homologous case: semantic transitivity

Heterologous case: human-in-the-loop **causal reasoning**

"AI surveillance → privacy infringement" + "Poor UX → user stress"

Composed threat: Unregulated AI surveillance may indirectly cause user stress

Compositionality - Oracle(s)

An **Oracle** is responsible for the *threat composition*

Low-Human-Intervention Oracle *(Fully Automated)*

Medium-Human-Intervention Oracle *(Semi-Automated)*

High-Human-Intervention Oracle *(Human-Driven)*

Multiple oracles can be used to *cross-validate* each other's outputs

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
SERICS SECURITY AND RIGHTS IN THE CYBERSPACE

Compositionality - Homologous Case

T_1: "Sessions enable unauthorised data access"

T_2: "Weak session authentication allows session hijacking"

$$0.9 = cos(emb(T_2), emb(T_1)) > \tau = 0.85$$

<u>Composed threat</u>:

"Weak session authentication → session hijacking → unauthorised data access."

Compositionality - Heterologous Case

T_a: "AI surveillance infringes on privacy rights"

T_β: "Poor UX leads to user frustration and non-compliance"

Human adds bridge: "users learn to bypass monitoring to avoid discomfort"

Composed threat:

"Unregulated AI surveillance infringes on privacy rights, prompting users to bypass monitoring, which leads to frustration and non-compliance"

Agenda

1. Introduction
2. Methodology
3. **Partial Validation**
4. Conclusions

SERICS

SECURITY AND RIGHTS IN THE CYBERSPACE

Proposed Validation Plan

Expert questionnaire: Cybersec practitioners, AI ethicists, HCI experts

Dimensions: *Applicability • Plausibility • Clarity • Relevance • Redundancy*

Format: Closed & open questions on composed threats

Validation Questionnaire

1. **Applicability** – *Does this threat apply to the domain "X" (e.g., financial services, public sector, healthcare)?*
2. **Plausibility** – *Is the causal link between the two components of the threat logically sound?*
3. **Clarity** – *Is the threat clearly formulated and understandable without additional context?*
4. **Relevance** – *Would you consider this threat relevant for inclusion in a cybersecurity risk assessment framework?*
5. **Redundancy** – *Does this threat overlap with any known threat categories or existing entries you are familiar with?*

Agenda

1. Introduction
2. Methodology
3. Partial Validation
4. **Conclusions**

SERICS

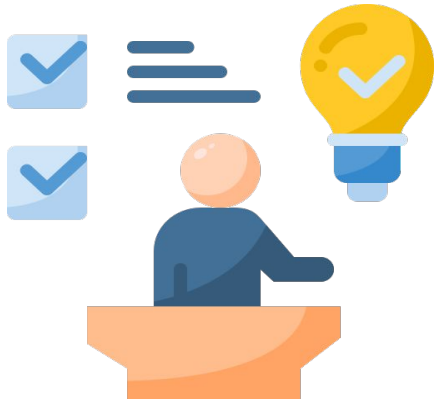
SECURITY AND RIGHTS IN THE CYBERSPACE

Conclusions

We advance a **HAI-powered threat elicitation methodology** to deal with the **compositional nature** of *heterogenous infrastructures*

Future work:

- Consolidation of the results on *AI, Cyber, Human Factor*
- **D5** — Prototype on *Web of Things & Industry 4.0*



Thanks for your attention!

For more information or questions:

nas.inf@studium.unict.it

https://nas.dmi.unict.it/