Towards Automated Threat Elicitation from the AI Act

S. Di Mauro, **M. Raciti**, G. Bella

SECAI 2025

IMT SCHOOL FOR ADVANCED STUDIES LUCCA

Università di Catania

26/09/25 – Toulouse, FR

Gaps and Contributions

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**

Our work:

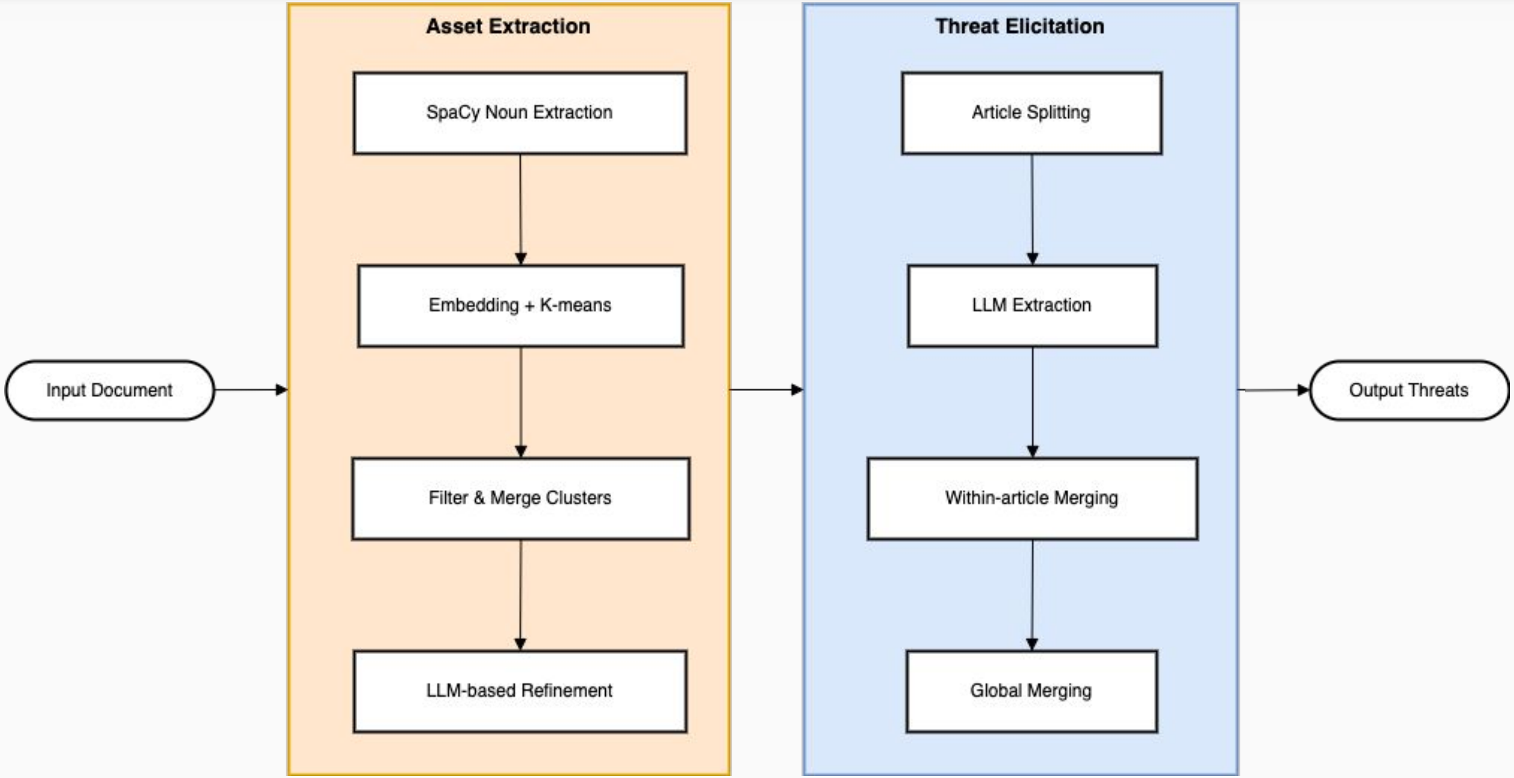
Takes a **Human Artificial Intelligence (HAI)** approach to automate *threat elicitation*

Applies such approach to the **AI Act**

Agenda

1. Introduction
2. **Methodology**
3. Application on AI Act
4. Validation
5. Conclusions

Methodology in a Nutshell



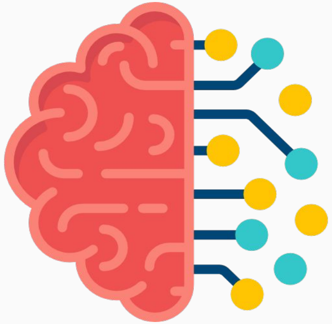
Agenda

1. Introduction
2. **Methodology → <u>Asset Extraction</u>**
3. Application on AI Act
4. Validation
5. Conclusions

Asset Extraction

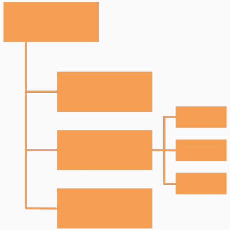
Based on *Natural Language Processing (NLP)* and *Clustering*, with an *LLM-based refinement*

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



Asset Extraction - NLP

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



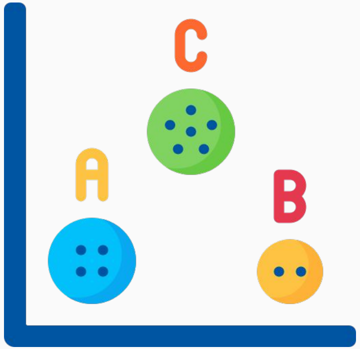
“Organizations shall ensure the integrity of personal data by implementing encryption and access controls.”

["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 → <u>Threat Elicitation</u>**
3. Application on AI Act
4. Validation
5. Conclusions

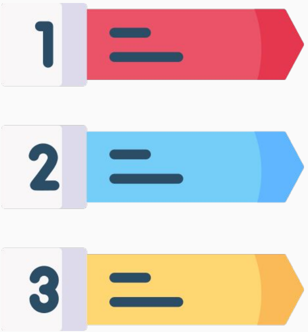
Threat Elicitation

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.

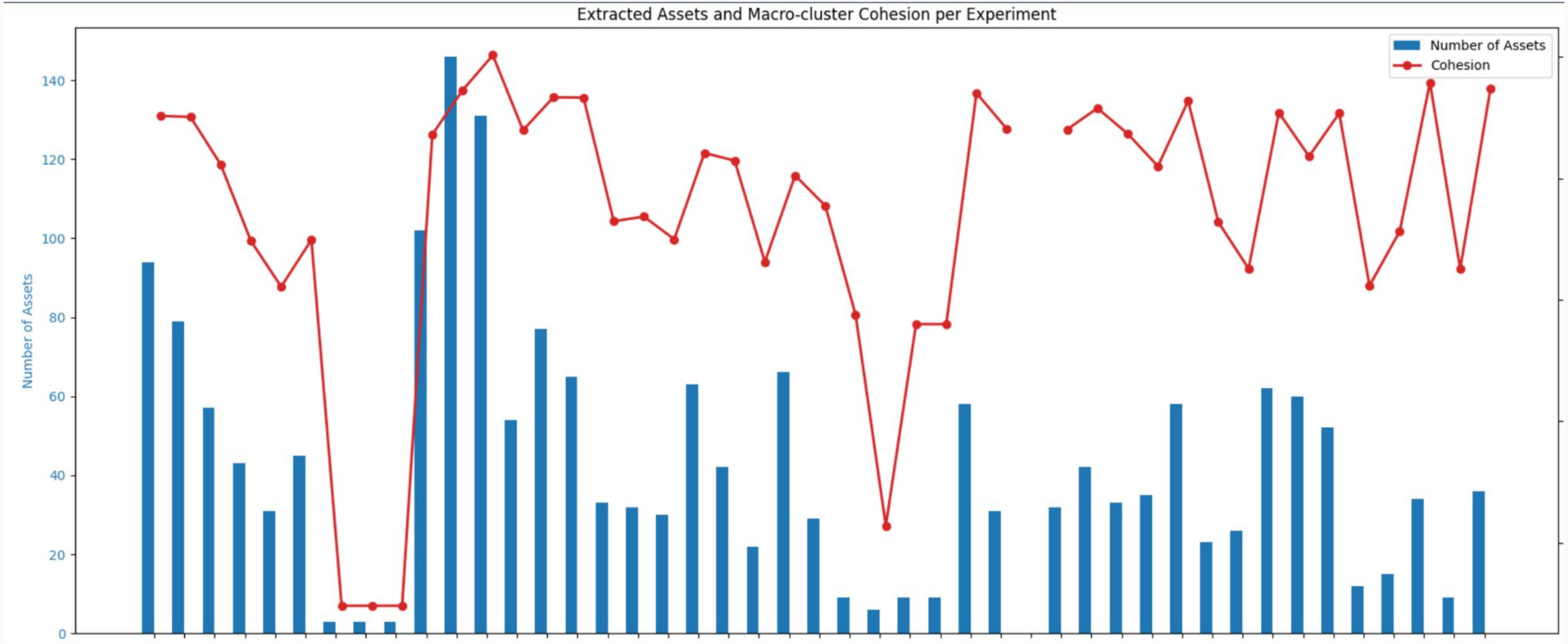
Agenda

Experimental Testing to Evaluate Thresholds

Table 1: Threshold values used in the experimental evaluation.

Threshold	Name	Values
Detection	`threshold_values`	$\{0.10, 0.15, 0.20, 0.25, 0.30\}$
Similarity	`similarity_threshold_values`	$\{0.50, 0.60, 0.70\}$
Merge similarity	`merge_similarity_threshold_values`	$\{0.70, 0.75, 0.80\}$

Performance Comparison of the Runs



Extracted Assets

Table 2: Sample of extracted assets from AI Act

Assets		
SYSTEMS	SYSTEM	SECURITY
MONITORING	DETECTION	CAMERAS
SURVEILLANCE	THREATS	CYBERSECURITY
ALARM	CYBER	CYBERATTACKS
BIOMETRICS	OPERATORS	SERVICES
ACCESS	PROVIDERS	NETWORK
INTERNET	MESSAGING	TELECOMMUNICATION
INSURANCE	MODELS	MODELLING
PROCESSING	DATA	STORING
SOFTWARE	OUTPUTS	ALGORITHMS

Extracted Threats

We extracted a total of
38 AI-related threats

Table 3: Sample of extracted threats from AI Act

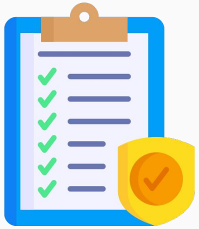
Threat	Explanation
AI Misuse that may cause harm, infringe on rights, or manipulate behaviors without proper regulation and oversight.	AI systems may be deployed in ways that cause harm, infringe on rights, or manipulate behaviors without proper regulation and oversight.
Discrimination and Bias arising from biased AI algorithms and datasets, resulting in discriminatory outcomes that violate fundamental rights and ethical standards.	Biases in AI algorithms and datasets can result in discriminatory outcomes, violating fundamental rights and ethical standards.
Lack of Transparency caused by opaque AI decision-making processes that hinder understanding, trust, and the ability to rectify AI behaviors, increasing misuse risks.	Opaque AI decision-making processes hinder the ability to understand, trust, and rectify AI behaviors, increasing the risk of misuse.
Privacy Violation from improper handling of personal and biometric data by AI systems, infringing on individuals' privacy rights and allowing data misuse.	Improper handling of personal and biometric data by AI systems can infringe on individuals' privacy rights and lead to data misuse.
Adversarial Attacks targeting AI systems with adversarial inputs designed to deceive or manipulate outputs, compromising reliability and security.	AI systems can be targeted by adversarial inputs designed to deceive or manipulate their outputs, compromising system reliability and security.

Beyond a Technical Catalogue

Privacy Violation resonates with **Articles 10–11** on data governance and quality, which demand lawful handling of personal and biometric data

Discrimination and Bias reflects **Article 10(3) and Recital 44**, mandating that datasets and outputs avoid discriminatory effects

Inadequate Human Oversight is directly addressed in **Article 14**, which requires that high-risk AI systems incorporate mechanisms for meaningful human control.



Agenda

1. Introduction
2. Methodology
3. Application on AI Act
4. **Validation**
5. Conclusions

Questionnaire Design

As **cybersecurity practitioners** and **AI ethicists**, we are all aware of these facts:

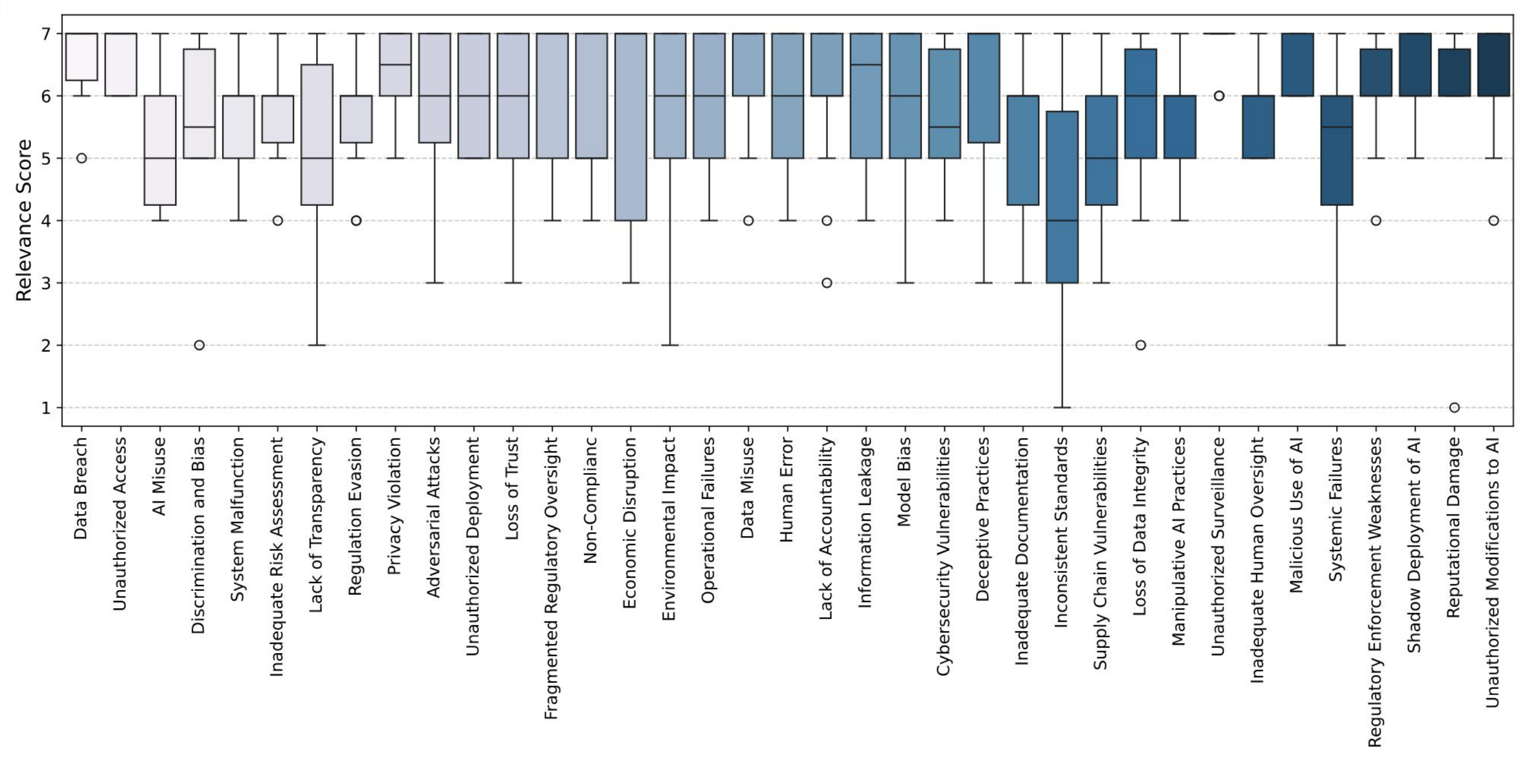
- Artificial Intelligence depends on large-scale data availability

- Big data enhances AI model precision, adaptability, and real-time processing

- The use of personal data in AI raises security, legal, and governance challenges

How relevant do you find the following threat with respects to those facts?

Validation Outcomes



Limitations

Thresholds may vary in other domains → *experimental runs*

LLMs are **non-deterministic** → *multiple runs to converge*

Generalisation is limited → *future work*

Limited set of responders → *future work*

Agenda

1. Introduction
2. Methodology
3. Application on AI Act
4. Validation
5. **Conclusions**

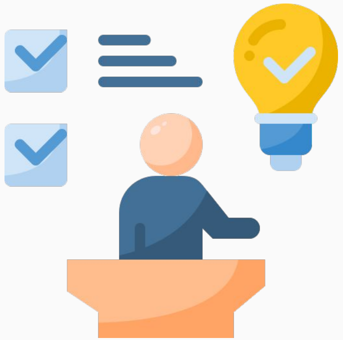
Conclusions

We advanced a **HAI-powered threat elicitation methodology** leveraging NLP and LLMs

We elicited a total of *38 AI-related threats* from the **AI Act**

Future work:

- Explore LLM fine-tuning and support to multi-lingual documents
- Extend the methodology to other regulatory frameworks
- Refine the validation through larger and more diverse expert panels



Thanks for your attention!

For more information or questions:

mario.raciti@imtlucca.it – mario.raciti@phd.unict.it

https://tsumarios.github.io/

@tsumarios

https://linkedin.com/in/marioraciti

Non-malicious QR (maybe)