International Journal of Information Security (2023) 22:1333–1354
https://doi.org/10.1007/s10207-023-00693-z

REGULAR CONTRIBUTION

AiCEF: an AI-assisted cyber exercise content generation framework using named entity recognition

Alexandros Zacharis[1] · **Constantinos Patsakis**[2,3]

Published online: 19 April 2023

Abstract
Content generation that is both relevant and up to date with the current threats of the target audience is a critical element in the success of any cyber security exercise (CSE). Through this work, we explore the results of applying machine learning techniques to unstructured information sources to generate structured CSE content. The corpus of our work is a large dataset of publicly available cyber security articles that have been used to predict future threats and to form the skeleton for new exercise scenarios. Machine learning techniques, like named entity recognition and topic extraction, have been utilised to structure the information based on a novel ontology we developed, named Cyber Exercise Scenario Ontology (CESO). Moreover, we used clustering with outliers to classify the generated extracted data into objects of our ontology. Graph comparison methodologies were used to match generated scenario fragments to known threat actors' tactics and help enrich the proposed scenario accordingly with the help of synthetic text generators. CESO has also been chosen as the prominent way to express both fragments and the final proposed scenario content by our AI-assisted Cyber Exercise Framework. Our methodology was assessed by providing a set of generated scenarios for evaluation to a group of experts to be used as part of a real-world awareness tabletop exercise.

Keywords Cyber security exercise scenario · Artificial intelligence · Cyber security exercise ontology

1 Introduction

Cyber security exercises (CSEs) are increasingly becoming an integral part of the cybersecurity training landscape [20], providing a hands-on experience to personnel of both public and private organisations worldwide. A CSE, as described in the ISO Guidelines for Exercises [18], is "*a process to train for, assess, practice, and improve performance in an organisation*". ENISA defines a CSE as "*a planned event during which an organisation simulates cyber-attacks or information security incidents or other types of disruptions to test the organisation's cyber capabilities, from being able to detect a security incident to the ability to respond appropriately and minimise any related impact.*" [7].

1.1 Problem setting and objectives

The creation of CSE content is a painstaking process that requires a deep understanding of the current threat landscape and the historical threats and incidents faced by an entity and the corresponding sector. Furthermore, training employees with simulated incidents is the closest method to testing the preparedness and effectiveness of measures and procedures set in place. Creating a relevant and dynamic content for developing CSE scenarios requires expertise and resources often lacking among most organisations.

The main objective of our work is automating the generation of structured CSE scenarios based on a pool of unstructured information with little experience in scenario building expected from the Exercise Planner (EP).

The standard method for preparing an exercise scenario [18] lays down three layers, namely events, incidents, and injects. After developing a scenario, an organisation must ensure that it contains only necessary information. More-

✉ Constantinos Patsakis
kpatsak@unipi.gr

Alexandros Zacharis
alexandros.zacharis@enisa.europa.eu

[1] European Union Agency for Cybersecurity (ENISA), Athens, Greece

[2] Department of Informatics, University of Piraeus, Karaoli & Dimitriou 80, 18534 Piraeus, Greece

[3] Athena Research Center, Marousi, Greece

over, it must be designed to test participants' capabilities in a stressful environment. Events, at the first level, provide the general description of an exercise scenario. Depending on previously decided objectives and aims, the number of events can differ from one exercise to another. Each event would have a specific set of consequences at the second level. These consequences are called incidents. An event can have multiple consequences, which can affect each other. On the third level, injects facilitate the communication of events and incidents to the exercise participants. An ideal inject would provide exercise information and problems to be solved. At the same time, it would indirectly force participants to act on those consequences and make decisions.

The proposed scenarios should satisfy the specifications provided by the EP. Such specifications can be the training topics and objectives, the sector to focus on or specific threats of interest that are currently or will be trending in the future. For simplicity, in what follows, when referring to sectors, we will refer to the ones of Directive (EU) 2022/2555 of the European Parliament and of the Council on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive) [12]; however, any other such classification can be used. More specifically, the objectives can be summarised as follows:

1. Create an ML-powered Exercise Generation Framework that would:
   (a) Generate structured exercise scenarios that reflect a sector's current or future threat landscape, including potential threat actors and the corresponding techniques, tactics, and procedures (TTPs).
   (b) Generate scripted events and incidents that could materialise in the context of a real attack against an organisation belonging to any NIS 2 defined Sector
   (c) Identify and describe artefacts that could accompany the exercise scenarios as potential injects
2. The generated scenarios should be expressed in a structured way or format, following an Ontology. The generated outputs should be both machine and human-readable.
3. The proposed methodology and tools created should provide qualitative and quantitative added value in CSE development and cyber-awareness by measuring the following Key Performance Indicators (KPIs):
   (a) Improve the speed in CSE generation (quantitative)
   (b) Improve quality in CSE generation (qualitative) for inexperienced EPs
   (c) Improve the relevance of proposed CSE scenarios to the current threat landscape (qualitative)

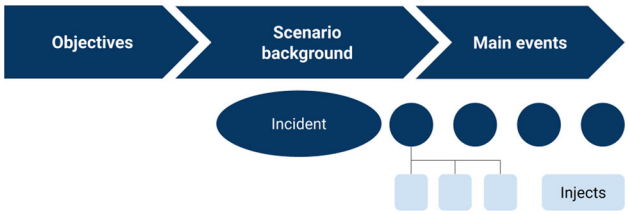


Fig. 1 Cyber exercise structure according to ISO 22398:2013 [18]

The use of case studies will help measure the results of the KPIs set by comparing the traditional exercise generation methods and tools versus the proposed ones through an evaluation provided by an Ad-hoc Cyber Awareness Expert Group [1] [2] that will peer review the outputs of the aforementioned methodology.

1.2 Main contributions

The contribution of this work is twofold. Initially, we can identify future cyber-attack trends in a specific sector and propose customised awareness training topics by clustering them accordingly. Then, we automate the process of generating the corresponding content for cyber awareness exercises with machine learning (ML).

Our proposed methodology, which a set of tools will accompany, allows an inexperienced EP to fully structure CSE scenarios from free text following our proposed Cyber Exercise Scenario Ontology (CESO). The exercise structure will follow the traditional Scenario—Events -Incidents—Injects tree structure ISO 22398:2013 [18] as depicted in Fig. 1. Additional cyber exercise content will be generated to complement the scenario and proposals for the fittest of a set of given training topics to better prepare an organisation for an imminent cyber crisis.

Through our work, we fill the gap in the lack of expertise by the average cyber security expert that acts as an Exercise Planner and provides the tools and the methodology to design CSE scenarios in an easy, automated, and structured way. To achieve this, we combine the power of machine learning (ML) and, more specifically, named entity recognition (NER) with a set of novel Cyber Exercise Scenario Ontology (CESO) and CSE scenario generation framework dubbed AiCEF. Finally, an evaluation methodology and its results are presented along with ideas for future work.

[1] https://www.enisa.europa.eu/topics/cybersecurity-education/ad-hoc-working-group-awareness-raising.

[2] "The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union Agency for Cybersecurity (ENISA). Neither the European Union institutions nor any person acting on their behalf may be held responsible for any use that may be made of the information contained therein.

2 Related work

CSEs, also known as Cyber Defense Exercises (CDX), have been considered an effective way to implement an engaging security awareness training [13, 42] experience. CSEs have been characterised as a highly effective method to provide an ultimate learning experience [3], helping individuals or teams of varying expertise improve a range of skills related to information security. Furthermore, via exercising, organisations can uncover gaps in security policies, procedures and resources [9, 16] leading to awareness training, tools and policy improvements.

Previous work in the CSE domain [40] has highlighted the use of cyber defence competitions or live-attack exercises as a very effective way of teaching information security [10, 19], helping teams design, implement, manage and defend a network of computers [1, 6, 7, 30, 31]. Vigna [46] and Mink [27] further support these findings.

Further research was conducted on cyber defence competitions [36, 49] and the most suited architecture [41] and tools and techniques to be used to create an active learning experience were described by Green et al. [15]. Patriciu and Furtuna [34] presented several steps and guidelines to be followed when designing a CSE. White [48] introduced a different approach to such live CSEs, presenting lessons learned and providing suggestions to help organisations run their own exercises. Other works in the literature examined how to run CSEs, using a service provider model [26].

CSEs can be used as a tool to generate scientifically valuable datasets for future security research [38, 43] and help uncover hidden risk from weak Security policies and/or procedures [37]. CSEs can even be used to measure performance against specific standards [11] or team effectiveness based on behavioural assessment techniques [14]. Moreover, experiments using various platforms like the RINSE simulator [23] or using realistic inter-domain routing experiment platform [22], for the rendering of network behaviour.

Focusing further on the human aspect, Job Performance Modelling (JPM), using vignettes for improving cybersecurity talent management through cyber defence competition design, was described by Tobey [44].

A successful CSE counts heavily on the use of a robust scenario. Exercise scenarios must describe worst-case scenarios that participants can relate to and are realistic enough to trigger seamless engagement. Intuitive scenarios can be a powerful tool that can predict future states or situations [3], [13]; incorporating issues to be resolved, interactions and consequences [15], [14] leading to a constructive training experience.

An exercise's scenario is a sequential, narrative account of a hypothetical incident that provides the catalyst for the exercise and is intended to introduce situations that will inspire responses and thus allow demonstration of the exercise objectives [41]. In the context of CSEs, a scenario defines the training environment that will lead participants towards fulfilling the exercise objectives [21] set. The cyber security problem described in a scenario itself portrays a structured representation named Master Scenario Events List (MSEL), which serves as the script for the execution of an exercise [41]. CSE scenarios formats can vary [35], but two are the most prevalent:

- Outlined scenarios: Provide a general summary of the impact of an event on assets. [39]
- Detailed scenarios: Contain exhaustive information sequentially describing the event's impact on specific services or sections of an organisation, along with a timeline for restoring key functions. [17]

Recent trends in attack recognition utilise AI, ML, and NLP tools and techniques to empower their efficiency. However, there needs to be a more dedicated methodology focusing on CSE scenario generation. There is a need for a methodologically built and annotated CE corpus that could train multiple algorithms for Cyber Exercise elements. Such a corpus should focus on the syntactic and semantic characteristics of the cyber exercise components and broaden our understanding of the malicious patterns used in cyber incidents that can be reused for CSE material. A similar approach to the one used in building and evaluating an annotated Corpus for automated Recognition attacks has been utilised [45], only this time to extract CSE relevant objects.

Following Cyber Security related ontology creation examples [33], ontology -based scenario modelling for CSEs have already been proposed [47]. Still, an ontology that is truly compatible with Machine Learning algorithms is missing and will be the focus of our work.

3 Cyber exercise scenario ontology (CESO)

Our work so far highlighted the need for a common CSE scenario ontology for translating the various parts of an exercise while keeping a close link to popular already used ontologies for cyber incident representations. The analysis of the domain revealed many taxonomies for different areas of the cybersecurity domain (types of attacks, vulnerabilities, sectors, harm) but those needed to be linked together in a model that allows for an EP to represent a CSE accurately.

To build our ontology, the following questions were raised:

1. What is the scope of the ontology?
2. Should we consider reusing existing ontologies or taxonomies?
3. What are the important terms in the ontology?

The scope of the ontology was determined by asking competency questions to experienced EPs that helped us identify the most important terms. A key priority was interoperability to ensure that the proposed ontology could be integrated with existing tools and frameworks. Moreover, the proposed ontology should describe a cyber security incident using popular cyber security frameworks. Finally, the ontology should be easily implementable using extraction via named entity recognition (NER) to allow the easy ingestion of online content.

We also used the domain expert's knowledge to identify prominent existing ontologies and ways to reuse them. The steps followed were:

1. Define the scope of our ontology,
2. Identify other ontologies or taxonomies that can be used/reused,
3. Define the main concepts and the relationships between them,
4. Define the properties of the concepts,
5. Implement the ontology.

3.1 Scope

The scope of the defined model was to target an efficient and robust way of representing cyber incidents in the context of a CSE. After all, a CSE is a collection of simulated incidents provided to players in an orchestrated way to achieve the exercise's objectives.

The exercise ontology presented is incident-centric, focusing on using a bottom-up approach that allows us to identify and describe incidents first so we can group them into Events and then cover the full generation of CSE scenarios that fit the high-level objectives set.

The first building blocks, incidents, are assigned injects and mitigation actions that match the expected scope of the scenario. Injection timing is configured on the attribute level of each object. As we build toward the higher level of the exercise, the scenario is formed. The selected format should allow for the scenario's portability to various existing tools (ex. MISP[3]) and support a decentralised type of CSE execution.

3.2 Ontologies/taxonomies to be (re)used

A set of existing ontologies, taxonomies, frameworks, standards, and formats have been explored with relevance to cyber security and a focus on the representations of the key element of CSEs from the point of their very building blocks being the incidents to be simulated. Our research concluded that a combination of the following would provide the necessary means: ISO 22398 [18], MITRE ATT&CK [29] and Cyber Kill Chain [24], MITRE CVE [28], and STIX 2.1 [32].

We chose STIX 2.1 as the basis for our ontology, which defines a taxonomy of cyber threat intelligence to be extended to cover our need to describe CSE scenarios. ISO 22398 best describes the structure of the cyber exercise components and was used to help us repurpose STIX 2.1 to cover our scope. The STIX 2.1 model describes an adversary and adversary activities in appropriate data structures by default. STIX Domain Objects cover: Threat Actor; Malware; Tools; Campaign; Intrusion Sets, and Attack Patterns (referencing the Common Attack Pattern Enumeration and Classification taxonomy, CAPEC), perfectly covering what is called incident and injects in the CSE nomenclature. STIX 2.1 supports by default the MITRE ATT&CK, MITRE CVE and Cyber Kill Chain frameworks, helping us achieve our goal for maximum interoperability. Moreover, intelligence (CTI) in a consistent and machine-readable manner, allowing security communities to understand better what computer-based attacks they are most likely to face and anticipate and/or respond to those attacks faster and more effectively.

This helps us build on top of these communities to reuse existing tools and share CSE scenarios represented in the very same format.

3.3 Scenario augmented model

Based on the bottom-up approach, a scenario augmented model (SAM) is proposed in two layers that cover both the informational and operational aspects with the same objects but utilise different attributes.

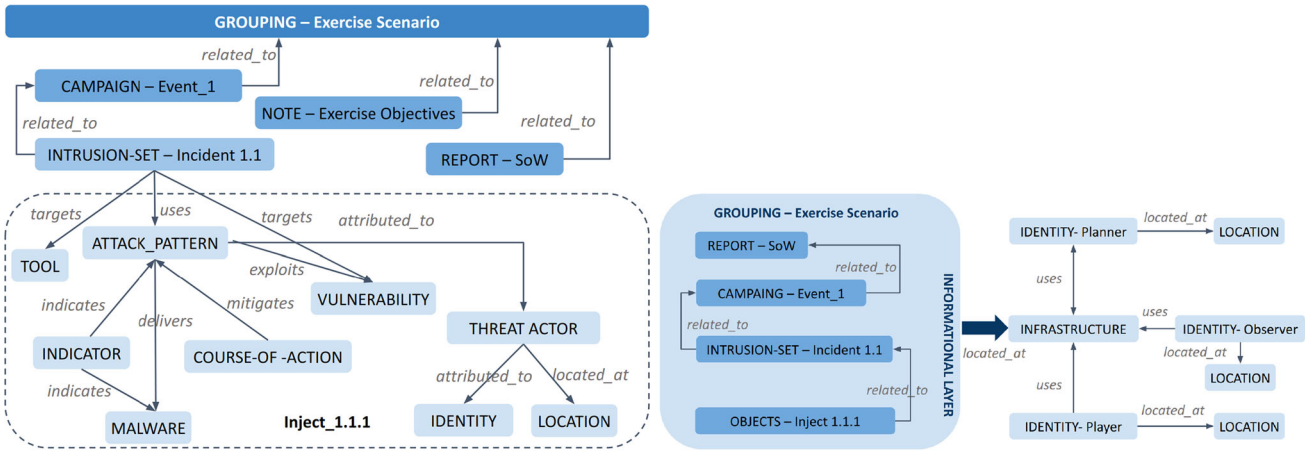
The informational layer covers the context and main attributes of scenarios. Figure 2a describes the key relationships in the informational layer.

The whole exercise is grouped using the grouping objects. The object holds information related to the exercise's name, description and scenario. All Events, Objectives and their matching objects (Campaign, Note, Report) are related to the Exercise Scenario along with the matching "State of the World (SoW)". The SoW includes details such as the status of various simulated systems and networks, the simulated geopolitical landscape, and any simulated incidents or events that have taken place.

One or more Incidents (Intrusion Set) can be related to Events. From there, various objects with interlinked dependencies form the Inject in a Course of Actions Instance that refers to all related objects of an Attack Pattern.

An Inject can contain the following objects: Attack Pattern, Tool, Vulnerability, Indicator, Malware Threat Actor (who is attributed and Identity and is located at a Location) and a Course Of Action. Injects do not have to be related to an Event or Incident. Such examples are the STARTEX or

[3] https://www.misp-project.org/.



(a) Informational Layer of CESO. (b) Operational Layer of CESO.

Fig. 2 The informational (left) and operational (right) layers of CESO

Table 1 CSE components to STIX 2.1 objects mapping

CSE Component	STIX 2.1 Object
Exercise details and scenario background	Grouping
Objectives	Note
Events	Campaign
Incidents	Intrusion Set
State of the World (SoW)	Report
Injects	Tool, vulnerability, threat-actor, identity, location, attack pattern, malware, indicator, course-of-action, observed data, malware analysis, report
Exercise Platform	Infrastructure
Exercise Participant	Identity, location

ENDEX,[4] which can be represented only with a Course of Action object but are directly related to the Scenario.

The Scenario Operational Layer describes an exercise scenario's execution flow, mainly dealing with injects delivery to the intended recipients. There are two major interrelated parts: (1) the events/injects, which describe the detailed activities of the scenario and expected actions from the participants, and (2) the Participants.

The whole scenario, including Events, Incidents, and Injects, is stored in an Infrastructure object, representing the Exercise Platform. This platform is used by EPs (Identity) to design and conduct the exercise, Observers, and Players to interact with the Scenario. All Participants are located in the same or different Locations. The Operational Layer is illustrated in Fig. 2b.

3.4 Implementing the ontology

Keeping the structure of CSE intact, the following STIX 2.1 Objects have been repurposed to fulfil our goal to represent the main CSE components successfully covering SAM along with matching relationships (Table 1).

Objects STIX 2.1 defined objects as per specifications.

Relationships All relationships are implemented as per STIX 2.1 relationship object specifications. The relationships in Table 2 (representing the edges of the graph) have been identified between key objects, but more can be used.

Object Extension STIX 2.1 objects, extended with additional attributes/properties to cover the needs of CESO, as shown in Table 3.

4 Automated generation of cybersecurity exercise scenarios

To create the envisioned ML-powered Exercise Generation Framework, we opted to use Python and develop a set of tools that would perform a set of individual tasks in the form of steps, which would help an EP, regardless of his/her experience, to create timely and targeted CSEs. Conceptually, we split the process into six steps. Namely, data collection, data

[4] The [START]ing and [END]ing [EX]ercise injects.

Table 2 Relationships matrix

Source Object	Destination object	Relationship
Campaign	Grouping	related_to
Note	Grouping	
Report	Grouping	
Intrusion-Set	Campaign	
Course-Of-Action	Grouping	
Intrusion-Set	Tool	targets
Intrusion-Set	Vulnerability	
Intrusion-Set	Attack-Pattern	uses
Identity	Infrastructure	
Attack-Pattern	Threat-Actor	attributed_to
Threat-Actor	Identity	
Identity	Location	located_at
Attack-Pattern	Malware	delivers
Attack-Pattern	Indicator	indicates
Indicator	Malware	
Attack-Pattern	Vulnerability	exploits
Course-Of-Action	Attack-Pattern	mitigates
Course-Of-Action	Vulnerability	

processing and mapping, trend prediction, incident generation, enhancement, and storyline generation. The proof of concept framework we developed is AiCEF, and its general outline is illustrated in Fig. 4.

Its main components that are relevant to the work presented in this paper are the following:

- CESO: The Cyber Exercise Scenario Ontology used to describe the various components of a CSE
- AiCEF: The Cyber Exercise Framework used to model CSEs based on CESO with the use of Machine Learning
- MLCESO: The ML models trained to parse text and extract objects based on CESO
- IncGen: The incident generation module that models a CSE incident from the MLCESO extracted objects based on CESO
- CEGen: The cyber exercise generation module that models a CSE from the MLCESO extracted objects based on CESO
- KDb: A knowledge pool of incidents stored in a database. Extracted objects and other characteristics, including the STIX 2.1 blob, are stored in the database

To facilitate the reader, we map these components in a timeline diagram, see Fig. 3. This way, one can get a quick grasp of the role of each component in the flow and navigate through the rest of the sections understanding how these pieces fit in the greater picture.

The modular approach of AiCEF allows for customisation and local refinements and enables more interoperability. In the following paragraphs, we detail each component and then present the main steps to generate a concrete CSE scenario using AiCEF modules, providing some examples.

4.1 Machine learning to CESO (MLCESO)

The most important step in our methodology is the creation of the ML pipeline that will parse free text and extract objects in CESO, as defined in the previous section. To do so, we need to train our ML following a well-structured methodology consisting of three phases: Corpus Building, Corpus Annotation, and Corpus Evaluation using NER, which we detail below.

4.1.1 Corpus building

As shown in Table 4, four Incident Sources have been identified as the initial input to our corpus. All these websites cover a wide variety of cyber security incidents in article format that date many years. For simplicity, in this work, we collected incidents from 2020–01 till 2022–03, which accounts for 2000 articles. All relevant articles were collected through automated web scraping.

Then, the raw text was processed using Natural Language Processing (NLP) techniques to form a reduced Incidents Corpus (IC). Initially, all text was converted to the UTF-

Table 3 Objects extension matrix

Object	Attribute added	Type	Description
Course-of-action	Difficulty (optional)	Integer	An integer from 1 to 5 (1 being easy and five being hard) declaring how difficult a course of action is evaluated to be executed by the player to resolve an incident
Grouping	Scenario (mandatory)	String	A description that provides more details and context about the exercise scenario
Identity	Recipient_Group (optional)	String	The name of the recipient group in which players are split to receive different injects

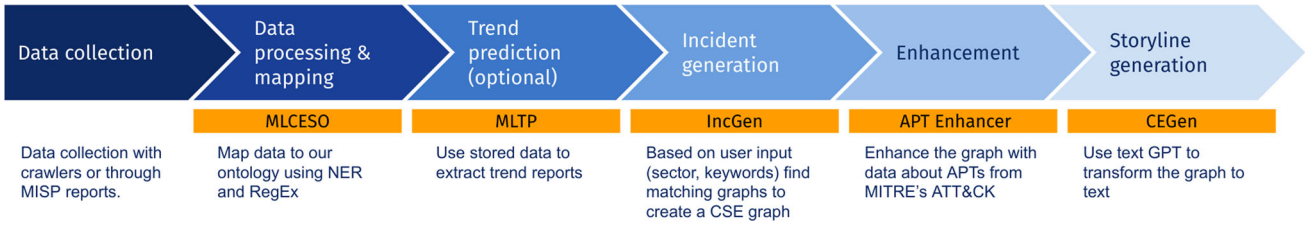


Fig. 3 Process flow and the corresponding modules of AiCEF

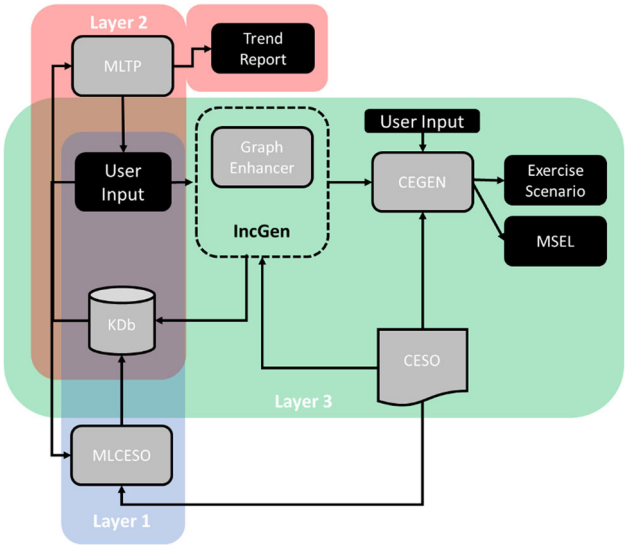


Fig. 4 High-level overview of AiCEF

Table 4 Corpus collection count

Webpage	Articles
`bleepingcomputer.com`	1000
`securityaffairs.co`	150
`zdnet.com`	350
`databreaches.net`	500
Total	2000

8 encoding scheme. Using dictionaries and the Textblob library,[5] we performed spelling corrections and removed special characters. Empty lines, specific stopwords and specific punctuation marks were removed using traditional NLP libraries like NLTK[6] and spaCy. [7] Moreover, all HTML or other programming codes, URLs, and paths were removed. Any illegal characters were also stripped, and all text was transformed to lowercase.

The standard Penn Treebank [25] tokenisation rules were utilised for sentence tokenisation, and finally, standardisation processes were applied to tune the Incidents Text to facilitate annotation. At the end of this step, a corpus composed of Incidents was formed. As discussed, the corpus, from now on referred to as IC, contains 2000 cyber security articles. This accounts for 35.745 sentences containing 819.690 words leading to a vocabulary of 24,594 terms. An example of a corpus line ready for annotation is the following:

```
{"text":"revil sodinokibi
   ransomware targets chinese
   users with dhl spam"}
```

4.1.2 Corpus annotation

Following the CESO ontology, a simple model was developed comprising six steps to represent the annotation task. Entities and interconnections were formally described to align the efforts of converting words to tags in an Annotators Reference Document (ARD). This file, along with the corpus guidelines and CESO ontology, was given to the annotators to perform the annotation task using Prodigy.[8] After completing the annotation, an inter-annotator agreement assessment took place using Cohen's Kappa metric, and the gold standard version of the IC was finally produced.

Our annotation methodology consists of the following steps.

Step 1: Setting the Annotation Objectives The main annotation objective was to create the appropriate semantic target to facilitate IC recognition by assigning the correct tag to in-context words in a sentence. Labelling all related words or sequences of words or text spans in the Cyber Incident context was crucial to perform efficient NER or text classification later. Each word or text span was labelled with a type identifier (tag) drawn from a vocabulary created based on the CESO ontology. It indicated what various terms denote in the context of a Cyber Incident and how they interconnect between them.

Our objective is to identify keywords, syntax, and semantic characteristics to detect i) threat actors, ii) cyber security

[5] https://github.com/sloria/TextBlob.

[6] https://www.nltk.org/.

[7] https://spacy.io/.

[8] https://prodi.gy.

Table 5 Annotation tags per category

Category	Tag	Link to CESO and STIX 2.1
Attacker	ATTACKER_TYPE	Threat Actor Attribute
	ATTACKER_NAME	Threat Actor Attribute, Identity
	ATTACKER_ORIGIN	Location
Attack	MALWARE_TYPE	Malware Attribute
	MALWARE_NAME	Malware Attribute
	ATTACK_TYPE (TECHNIQUE)	Attack Pattern
	VULNERABILITY	Vulnerability
Victim	SECTOR	Identity Attribute, Scenario
	ASSETS	Threat Actor Attribute
	TECHNOLOGY	Tool

Table 6 Annotation tags per category example

Category	Tag	Annotator A-Tags	Annotator B: Tags
TEXT	qbot malware dropped via context aware phishing campaign infects the energy sector.		
	russian hacking group claims 1000 windows machines compromised.		
Attacker	ATTACKER_TYPE	Hacking group	
	ATTACKER_NAME		
	ATTACKER_ORIGIN	Russian	Russian
Attack	MALWARE_TYPE	malware	malware
	MALWARE_NAME	qbot	qbot
	ATTACK_TYPE (TECHNIQUE)	phishing campaign	phishing
	VULNERABILITY		
Victim	SECTOR	energy	energy
	ASSETS	windows machines	windows machines
	TECHNOLOGY	windows	windows

incidents, and iii) victim characteristics, to tag them accordingly.

Step 2: Specifications Definition A concrete representation of the Annotation model to be used is created based on CESO.

An abstract model that practically represented the annotation objectives was defined. A three-category classification (Attacker, Attack, Victim) was introduced as the basis of this abstract model for identifying cyber-incident related terms in the text analysed. The category *other* represents all remaining words out of context.

Our model M consists of a vocabulary of terms T, the relations between these terms R, and their interpretation I. Thus, our model can be represented as $M =< T, R, I >$ where:

- T={CESO, Attacker, Attack, Victim, Other}
- R={CESO:: = Attacker|Attack|Victim|Other}
- I={Attacker= "list of attacker related terms in vocabulary", Attack ="list of Cyber Security Incident or Attack terms in vocabulary",
  Victim = {"list of victim-related terms in vocabulary"}
  Other = {"Other terms not related to the attacks"}}

Step 3: Annotator Reference To help annotators in element identification and element association with the appropriate tags, we provided them with documentation containing the tags in Table 5, which have been identified and mapped accordingly.

Step 4: Annotation Task the annotation process is performed

The annotation task aimed to label the words of the IC corpus based on their semantic and syntactic characteristics. Two cybersecurity experts were assigned to label the words based on their semantic characteristics. By annotating the semantic characteristics of the words with Prodigy, the context of each sentence was translated into CESO. Table 6 presents the annotation in action through some examples.

Table 7 Consistency matrix

Annotator		B				Total
	Category	Attacker	Attack	Victim	Other	
A	Attacker	397	10	4	24	435
	Attack	13	1722	8	9	1752
	Victim	10	2	926	15	953
	Other	16	10	12	21, 416	21,454
Total		436	1744	950	21,464	24,594

Table 8 AI models' scores

Category	Tag	Presicion	Recall	F1
Attacker	ATTACKER_TYPE	100.00	83.33	90.11
	ATTACKER_NAME	95.29	87.10	91.01
	ATTACKER_ORIGIN	Used Native Spacy LOC tag (no training)		
Attacker	MALWARE_TYPE	80.56	76.32	78.38
	MALWARE_NAME	95.29	87.10	91.01
	ATTACK_TYPE (TECHNIQUE)	88.60	87.07	87.83
	VULNERABILITY	87.50	84.00	85.71
Victim	SECTOR	85.84	84.07	84.95
	ASSETS	87.02	89.06	88.03
	TECHNOLOGY	87.60	89.93	88.70

Step 5-Golden Standard Creation: the final version of the annotated Incident corpus is generated.

The inter-annotator agreement (IAA) was validated using Cohen's Kappa [8]. The formula used is defined as follows:

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

where p_0 expresses the relative observed agreement, and p_e is the hypothetical probability of chance agreement.

The produced IC corpus has $N = 24594$ terms and $m = 4$ categories, and both annotators (A and B) agreed for the Attacker category 397 times, for the Attack category 1722 times, for the Victim 932 times and for the Irrelevant 21416.

Table 7 shows the contingency matrix where each x_{ij} represents the multitude of terms that annotator A classified in category i, but Annotator B is classified in category j, with $i, j \in \{1, 2, 3, 4\}$. The proportions on the diagonal (x_{ii}) represent the proportion of terms in each category for which the two annotators agreed on the assignment.

The observed agreement p_o is:

$$p_o = \frac{397 + 1722 + 926 + 21416}{24594} = 0,996$$

and the expected change agreement; thus, the proportion of terms which would be expected to agree by chance is:

$$p_e = \frac{\frac{436\times435}{24594} + \frac{1744\times1752}{24594} + \frac{950\times953}{24594} + \frac{21464\times21454}{24594}}{24594}$$

$$= 0,768(76,8\%)$$

so, according to Eq. 1 the Cohen's Kappa is $k = \frac{p_0 - p_e}{1 - p_e} = \frac{0,228}{0,232} = 0,98$. Thus, based the Cohen's kappa value of 0.98, we can safely conclude [50] that the level of agreement for the corpus annotation task was almost perfect.

4.1.3 Training and evaluation using NER

The following methodology has been used to train and evaluate our Named Entity Recognition (NER) agent.

1. **Preprocessing** The corpus has already been annotated, with each line of the corpus stored as a list of token-tag pairs. Each token was represented by a word embedding using the pre-trained English language model of the spaCy NLP library.
2. **Build a model using spaCy**
3. **Training** Training was conducted in spaCy by specifying a loss function to measure the prediction error and a batch-wise gradient descent algorithm for optimisation. One NER model was created per object. To improve accuracy, several iterations were conducted by expanding the annotation and retraining the model until an F1 Score of $\approx$ %80% was reached.
   One NER model was trained per object as presented in Table 8.

Table 9 AI models' scores vs reviewers evaluation. H: Hit, P: Partial, M: Miss

Category	TAG	Reviewer 1			Reviewer 2			Average			F1
	SCORE TYPE	H	P	M	H	P	M	H	P	M	
Attacker	ATTACKER_TYPE	90	5	5	90	5	5	90	5	5	90.11
	ATTACKER_NAME	63	20	17	59	24	17	61	22	17	91.01
	ATTACKER_ORIGIN	70	25	15	65	30	15	67.5	17.5	15	98
	MALWARE_TYPE	82	11	7	80	12	8	81	11.5	7.5	78.38
Attack	MALWARE_NAME	72	14	24	71	14	25	71.5	14	24.5	91.01
	ATTACK_TYPE (TECHNIQUE)	84	11	5	86	9	5	85	10	5	87.83
	VULNERABILITY	75	10	15	79	10	15	77	10	15	85.71
Victim	SECTOR	84	16	0	86	13	1	85	14.5	0.5	84.95
	ASSETS	90	7	3	90	8	2	90	7.5	2.5	88.03
	TECHNOLOGY	90	8	2	86	12	2	88	10	2	88.70

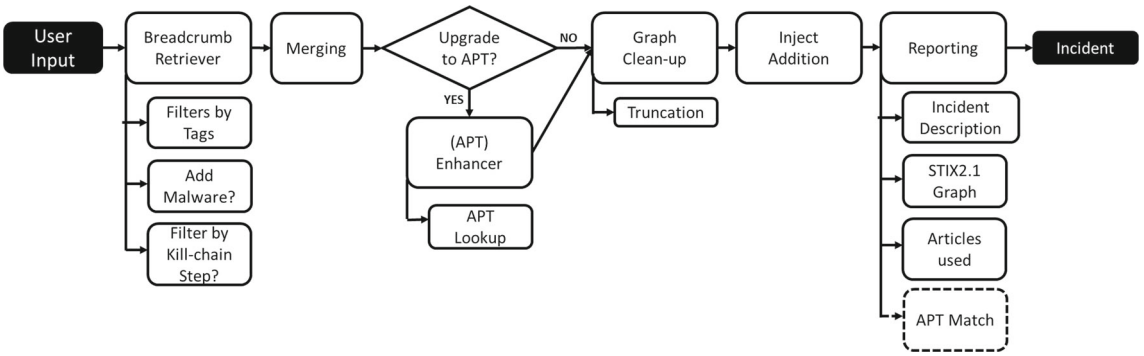


Fig. 5 The workflow of IncGen

4. **Evaluation** The performance assessment of the model was conducted by applying the model to the preprocessed validation data.

While the results seem satisfactory, one can achieve further performance improvements in some tags.

We made an extra evaluation step with two experts against a set of 100 articles not used before in the training or evaluation steps. The aim was to evaluate the models against the selected tags empirically. The two reviewers have scored the NER accuracy per tag as presented in Table 9:

- **HIT** The tag was correctly assigned or not.
- **PARTIAL** The tag was correctly assigned or not, but not for all values
- **MISS** The tag was either assigned wrongly or was not assigned at all when it should

The following findings should be highlighted:

1. The hit rate of four (4) NER models has been identified as very weak, with an abnormal difference from the F1 score identified in the previous step.
2. Names of Attackers or Malware can be a very vague topic to tackle using NER.
3. The Attacker's Origin cannot be properly identified with the use of the out-of-the-box SpaCy LOC NER model. Locations are identified but can be related to the victim or are irrelevant to the attacker's origin.
4. The vulnerability NER model misses the correct formatting of CVE. This issue can be solved using a regex that accurately detects CVE in the text in combination with the model generated.

4.2 Incident generation and enhancement (IncGen)

Incident creation is the most important step of the scenario generation procedure and consists of several steps to achieve maximum customisation (Fig. 5). All of the steps can be automated, generating a variety of Incidents from which a Planner can choose to fit most.

The EP can choose to provide specific text or articles for conversion to Incidents or rely on a dynamic generation based on filtering parameters and a search of the existing database. Incidents can be enhanced with activity simulating TTPs of known APT actors.

Table 10 Knowledge DB content per source

Source	Count
bleepingcomputer.com	1368
securityaffairs.co	169
zdnet.com	495
databreaches.net	938
Total	2970

To generate scenarios, a set of texts was used as a baseline and parsed to map with CESO for processing. The sources in Table 10 were utilised to create the knowledge database (KDb). To ensure relevance, a threshold system *maturity* was introduced to evaluate the maturity of the parsed articles and NER extracted tags. The scoring system, ranging from 0 to 185, is shown in Algorithm 1. In the implementation, a threshold of 50 was set to consider a text relevant for representing a standalone incident in AiCEF.

Algorithm 1 Compute the maturity of a parsed text.

Require: Set of Tags T
$maturity \leftarrow 0$
if $Attacker\_Type \in T$ OR $Attack\_Type \in T$ **then**
$maturity \leftarrow 50$
if $Vulnerability \in T$ **then** $maturity \leftarrow maturity + 15$
else $maturity \leftarrow maturity - 10$
end if
if $Malware \in T$ **then** $maturity \leftarrow maturity + 15$
else $maturity \leftarrow maturity - 10$
end if
if $Attack\_Type \in T$ **then**
$maturity \leftarrow maturity + 15$
if $Attacker\_Type \in T$ **then**
$maturity \leftarrow maturity + 50$
if $Technology \in T$ **then** $maturity \leftarrow maturity + 10$
end if
if $Sector \in T$ **then** $maturity \leftarrow maturity + 10$
end if
if $Assets \in T$ **then** $maturity \leftarrow maturity + 10$
end if
if $Attackers\_Origin \in T$ **then** $maturity \leftarrow maturity + 10$
end if
end if
end if
end if
return $maturity$

Two types of enhancements were applied to improve the automatically NER exported tags, namely Regular Expression (REGEX), which is a sequence of characters that defines a search pattern, and Hard-coded groups of Strings. Thus, the following tags have been further enhanced:

- Attackers Name: NER + Hardcoded Groups of Strings from MITRE APT list [29]
- Attackers Origin: No NER, Hardcoded Groups of Strings,
- Malware Name: NER + Hardcoded Groups of Strings from MITRE APT list,
- Technique: NER + Hardcoded Groups of Strings from MITRE APT list,
- Vulnerability: NER + CVE REGEX.

The above enhancements greatly improved the tag detection rates, achieving almost 99% in the Vulnerability tag. Moreover, based on the analysis of the most prominent extracted tags, the tag groups of Table 11 were assigned to the training topics meta tag to help categorise text for later use in an exercise scenario-building process. An output report and visualisation (using stixview[9] library) of IncGen utilising the improved MLCESO tag detection can be seen in Fig. 6.

4.3 APT enhancer

To simulate the activity of APT groups, a STIX 2.1 structure was created for each actor using the Groups from MITRE. Attributes and TTPs were automatically extracted to populate the database, generating a STIX 2.1 graph for comparison and enhancement purposes. During incident enhancement, the extracted graph is compared to known APT actors and the most similar is proposed for enhancement. The similarity score, based on a set of weighted properties and ranging from 0 to 100, is calculated using the STIX 2.1 Python API. In AiCEF, the EP can completely or partially merge the draft incident graph with that of known APT actors.

4.4 Storyline text generation

The Storyline Text Generator (STG) creates synthetic text based on predefined input. Using a Python text generator and Generative Pre-trained Transformer 2 (GPT-2),[10] AI large-scale unsupervised language model, which can create coherent paragraphs of text from small pieces of text input.

4.5 Trend prediction module (MLTP)

The trend prediction module provides EP with valuable information by analysing the KDb and extracting trends based on predetermined training objectives to generate a trend report. The MLTP process consists of three steps:

1. Receiving input such as Filter Tags
2. Extracting incident statistics based on specified sector and Training Objective

[9] https://github.com/traut/stixview.

[10] https://openai.com/blog/better-language-models/.

Table 11 Training topics

Training topic	Tags
INCIDENT HANDLING	MALWARE, RANSOMWARE, APT, CYBER, ATTACK, WEBSITE, HACKER, EXPLOIT, ZERO-DAY
GDPR	PRIVACY, DATA LEAKAGE, PERSONAL DATA, EXFILTRATION, CLOUD, SENSITIVE DATA, DATA, GOOGLE DRIVE, AWS, MEDICAL DATA, PASSPORT
CYBER HYGIENE	PASSWORD, ACCOUNT, USERNAME, LOGIN, ACCOUNTS, FILES, CREDENTIALS
PHISHING and SOCIAL ENGINEERING	PHISHING, SCAM, FRAUD, VISHING, IMPERSONATION, BEC, EMAIL, GMAIL
SOCIAL MEDIA	FACEBOOK, TWITTER, LINKEDIN, META, INSTAGRAM
BYOD	MOBILE, ANDROID, IOS, LAPTOP, IOT, GOOGLE PLAY

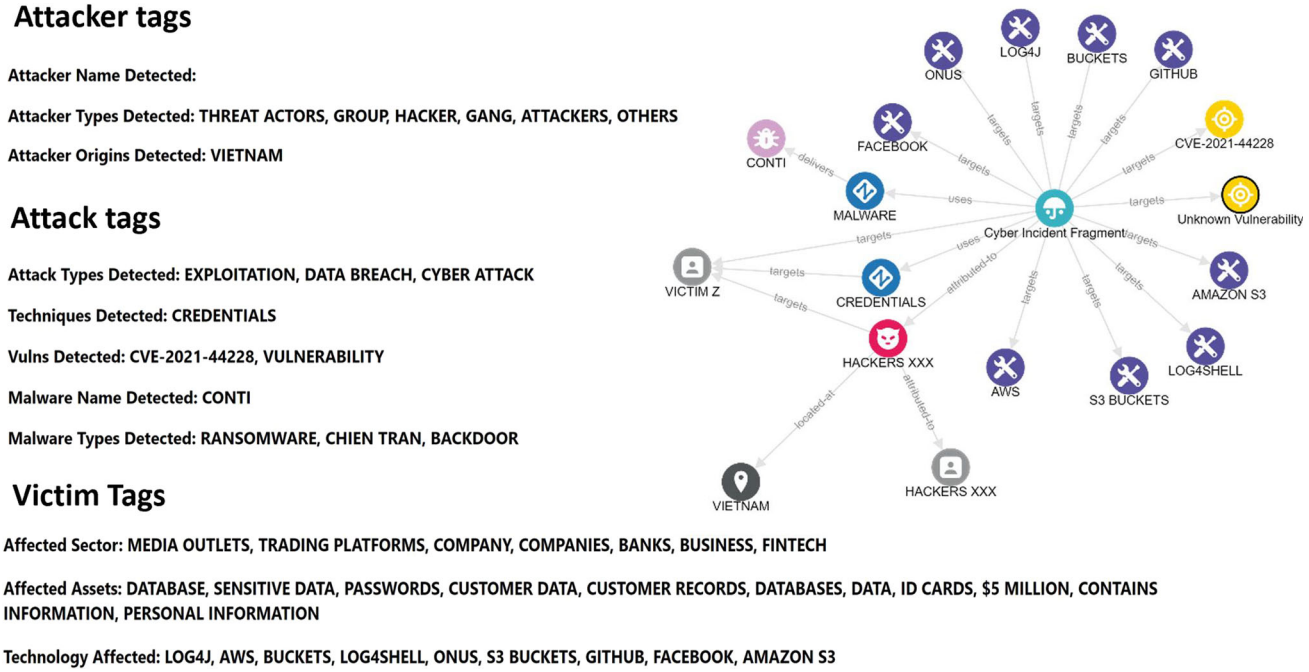


Fig. 6 IncGen output report and visualisation

3. Performing time-series analysis to plot and calculate future trends for a specific Attack Type and/or Training Objective.

In our implementation, we chose the SARIMA[11] equation to represent the trends on the existing KDb of 2970 articles as represented in Table 10. However, in future work, we intend to investigate further methods to boost the capabilities of MLTP, including the identification of micro-trends as the existing results are very promising [2].

[11] SARIMA is Seasonal ARIMA, or simply put, ARIMA with a seasonal component. ARIMA is a statistical analysis model that uses time-series data to predict future trends.

4.6 Putting everything together

Let us summarise the use of AiCEF and its modules with an example. An EP populates the Knowledge database (KDb) with incidents of interest, which are then converted into graphs based on the CESO ontology. When the EP wishes to create a new scenario for a cyber security exercise, they provide AiCEF with a set of keywords. To assist the planning process, AiCEF can generate a trend report that identifies trends relevant to the objectives at the time of the exercise execution. Based on the keywords, AiCEF crawls its database for the most relevant articles and returns a corresponding graph. The EP can then enhance the graph by merging it with that of known threat groups and filtering the graph according to the intended Cyber Kill Chain phases to be simulated. The resulting incident graph representation is then ready to be

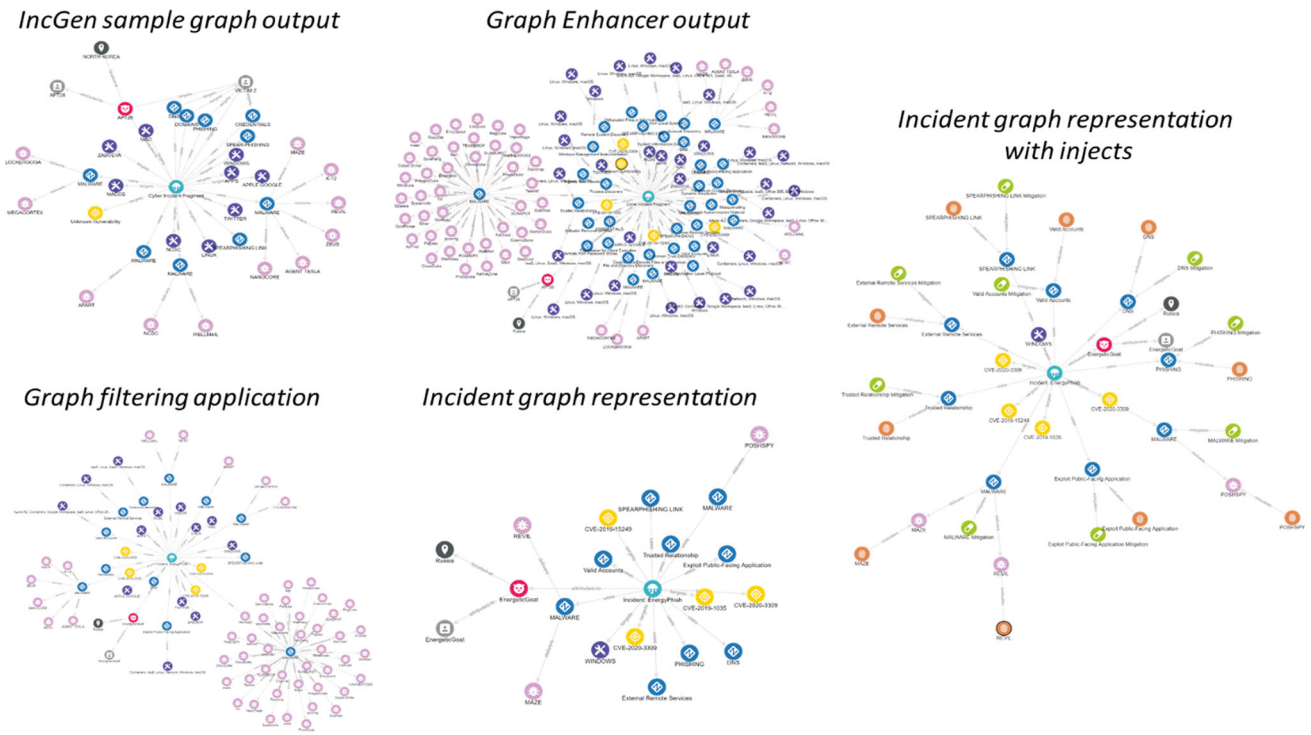


Fig. 7 IncGen execution flow with intermediate representation steps

populated with injects. A representation of the progress of an incident graph generation can be visualised in Fig. 7.

This process is repeated multiple times to generate the number of wanted incidents for a specific CSE. The EP follows the CEGen flow to compile a full exercise and generate a scenario (Fig. 8) and Exercise graph (Fig. 9).

5 Evaluation methodology and results

We developed a case study to help measure the effectiveness of our proposed framework and underlying methodology. To this end, the steps below were followed.

1. **Scenario Content Generation** A group of exercise planners of varying expertise have been used to generate the same exercise scenario using traditional exercise means and the AiCEF methodology and tools while being monitored on timeliness, effectiveness, creativity and methodology used.
2. **Content Evaluation** The reports were anonymised and given to a group of evaluators to grade the complexity, technical depth and richness of lessons learnt on the generated subset of exercise scenarios as per Objectives and KPIs set through a questionnaire.
3. **Results collection and Analysis** The results of this process were evaluated against the previously set KPIs to estimate:
   (a) Improved speed in Cyber Exercise Scenario generation (quantitative) using AiCEF.
   (b) Improvement in quality in Cyber Exercise Scenario generation (qualitative) for inexperienced Planners using AiCEF.
   (c) Improved relevance of proposed Cyber Exercise Scenarios to the current threat landscape (qualitative) using AiCEF.

5.1 Scenario content generation

Four EPs were selected to individually generate a CSE scenario according to the provided high-level exercise requirements and specifications, see Fig. 10. The EPs were split into two groups based on their previous experience with the task. All EPs have deep knowledge of cyber security, and their skill sets resemble that of a CISO.

Both groups consisted of one experienced and one inexperienced planner. The first group was briefly introduced to the basics of developing CSE scenarios, while the second one was provided with a course on using AiCEF and the accompanying tools. Both groups were provided with the same Scenario Template (ST) to fill in as an output of their task.

Then, we created a third group, consisting of Scripted Exercise Planner (SEP), using different parameters and flows of the AiCEF methodology and toolset.

The provided ST had the following generic structure:

State of the World (SoW)

Description: **STG Input**:: '*State sponsored Threat actor EnergeticGoat has historically targeted government organizations, non-government organizations (NGOs), think tanks, military, IT service providers, health technology and research, and telecommunications providers. With this latest attack, EnergeticGoat attempted to target the Energy Sector*'
STG Output: "State sponsored Threat actor EnergeticGoat has historically targeted government organizations, non-government organizations (NGOs), think tanks, military, IT service providers, health technology and research, and telecommunications providers. With this latest attack, EnergeticGoat attempted to target the Energy Sector in South Europe. More specifically the following targets have been identified: 1) EU energy firms that have received investments from the Energysubsidiary, 2) SA government units and departments who receive funding from the Energysubsidiary, 3) the SA energy regulator, 4) the Energy Sector's main lobbying and public relations organization, 5) the Energy Sector's largest customer, the Energysubsidiary, and 6) the Energy Sector's corporate regulator, the Energy Regulator of SA (ERA). Energy Sector entities and their customers and regulators are usually among the least likely organizations to be attacked by EnergeticGoat.
For example, the Energy Regulator of SA (ERA) is not a high-profile organization. It is a largely unregulated regulatory body that provides general oversight of the energy sector and serves as a key regulator and enforcement body of relevant energy laws in EU. Another example would be the Energy Corporation (Zesa). Zesa is a state-funded enterprise that provides essential infrastructure and related services to its citizens, including the electricity, gas, and broadband services that power large swaths in EU.
The attackers may have targeted other Energy Sector actors, such as Eskom or the SA Energy Institute (SAEI). However, it is thought that EnergeticGoat is specifically targeting Zesa and other organizations as a result of the threat actor's perceived inability to target the SA Energy Sector as a whole. EnergeticGoat's choice to target a South African energy firm with its latest attack is indicative of this perception."

Scenario Summary

EXERCISE NAME: Energy Test
EXERCISE OBJECTIVES: Test detection of Attacks and Incident Handling Procedures
EVENT 1: EnergyWipeOut ### Description: **STG Input**:: '*State sponsored Threat Actor EnergeticGoat is believed to be behind cyber-attacks against Energy Corp. The attackers used PHISHING & DATA DESTRUCTION techniques to*'
STG Output: "State sponsored Threat Actor EnergeticGoat is believed to be behind cyber-attacks against Energy Corp. The attackers used PHISHING & DATA DESTRUCTION techniques to infect the system with malware. In addition to the malware used, the attackers also used other tools to get around the security countermeasures of ENERGYCORP. It is understood that it will take at least a few days for the authorities to identify the attackers as they have only just informed the company about the attack. A spokesperson for the company said the company did not have any details as to the motive behind the attack as it was still investigating. "We are working with law enforcement to bring this to the most serious level. The situation is under investigation and we cannot comment any further than that"
*** Incident 1.1: EnergyPhish *** Description: **STG Input**:: '*EnergeticGoat deployed Phishing attacks against Energy Corp. Malicious word documents were sent*'
STG Output:"EnergeticGoat deployed Phishing attacks against Energy Corp. Malicious word documents were sent to Energy Corp. employees in an attempt to obtain personal information. On October 26, 2022, a new employee of Energy Corp., received a suspicious email containing a malicious word document. The email identified as 'malware' was in fact a phishing attack with the subject 'Vendors of Energy Corp. under Federal Investigation"

Fig. 8 Sample text of an AI-generated exercise

- Section 1: Storyline (SoW)
- Section 2: Scenario and MSEL
- Section 3: Scenario Analysis
- Section 4: Resources Used

We provided detailed instructions on the expected content per paragraph to all involved planners to streamline the information of the generated reports and create homogeneous outputs to be evaluated in the later step.

As a result, five complete exercise scenarios were generated, as shown in Table 12.

5.2 Scenario content evaluation

To evaluate the scenarios above, we conducted an anonymous online survey from 01/09/2022 to 30/09/2022. To avoid bias, we invited a number of evaluators from different cyber awareness and cyber exercise groups with varying expertise, ethnicity, and focus sectors to participate in the evaluation process. More precisely, we invited the Ad-hoc Cyber Awareness Expert Group of ENISA. In total, 16 experts responded, whose demographic statistics are illustrated in Table 13. Given that we have a representation of 66% of the group, we believe that the sample is significant, as they are experts. Moreover, we highlight that their allocation has been made through independent criteria, not from us, but from an individual international organisation on cyber security such as ENISA, which avoids possible biases.

The survey was in the form of an online questionnaire consisting of 11 questions. Eight questions were used to evaluate the generated Scenarios, two to be used as Turing test to determine whether the AI used could be identified by humans and a set of complementary questions for demographic and future improvement purposes. All five scenarios were provided using only the "Eval_Tag" parameter for tracking purposes without providing additional information on the authors of the scenarios.

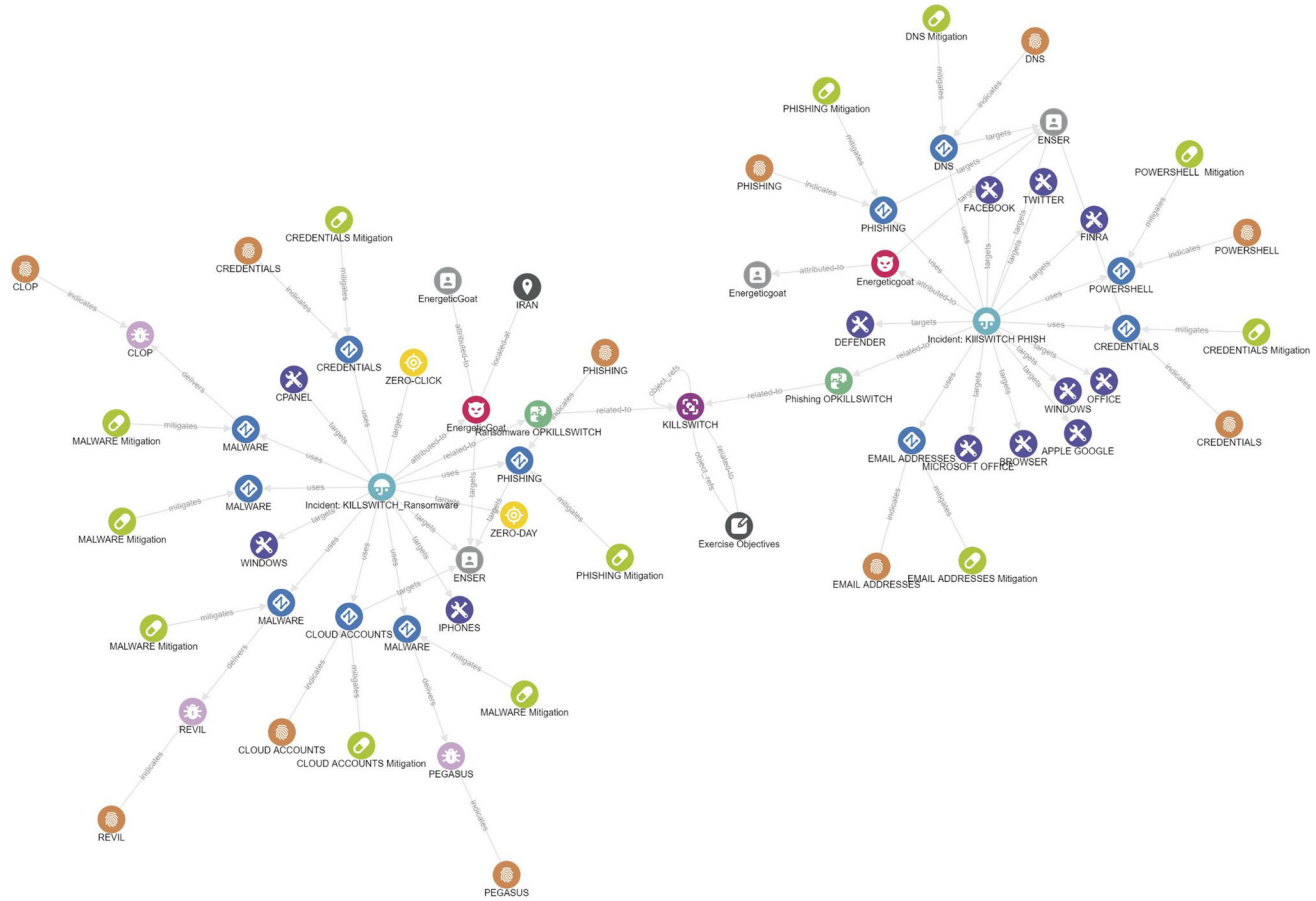


Fig. 9 Sample exercise graph visualisation

> Generate a CSE scenario for a cyber awareness exercise by filling in a provided Scenario Template. The CSE should include 2 events consisting of 1 incident each. All incidents should be accompanied by a short description of indicative injects to be sent to players. At least 3 inject descriptions per incident should be provided.
> The company which will use the scenario is an **Energy Service Provider** and all its Employees can be potential Players.
> The exercise should last between 2-4 hours and can include technical artefacts for analysis.
> The two main objectives of the exercise are:
>
> 1. Provide awareness to employees regarding Phishing Attacks
> 2. Provide awareness to employees regarding Ransomware Attacks
>
> These objectives can be updated and more can be added. The Scenario Development task will be timed and should not last more than 4 hours

Fig. 10 Task definition

The eight scenario evaluation questions and their corresponding scores in parenthesis were the following:

1. How do you evaluate the relevance of the State of the World text to the Objectives of the Exercise? (0–4)
2. How do you evaluate the relevance of the selected Events to the Objectives of the exercise? (0–4)
3. How do you evaluate the relevance of the selected Incidents to the Objectives of the exercise? (0–4)
4. How do you evaluate the Complexity of the Scenario? (0–1)
5. How do you evaluate the Technical Depth of the Scenario? (0–2)
6. How do you evaluate the Threat Actor's description? (1–3)
7. How do you evaluate the used resources? (0–2)
8. Would you be willing to use this Scenario based on the task description? (0–4)

To evaluate the use of AI for exercise content generation, we asked the expert the following questions:

1. How was the scenario generated?
2. How skilled was the planner?

Other questions revolved around the overall scenario development process:

Table 12 Details of the generated scenarios

Eval_Tag	Explanatory Name Tag	Exercise Expertise	AiCEF	Duration	Other Tools
ExSc1	Sc1:Exp(erienced)Hum(an)	YES	NO	2 h 00min	Google, MITRE
ExSc2	Sc2:Nov(ice)Hum(an)	NO	NO	2 h 35 min	Google
ExSc3	Sc3:Exp(erienced)Hum(an) and AI	YES	YES	1 h 20min	Online Resources
ExSc4	Sc4:Nov(ice)Hum(an) and AI	NO	YES	2 h 10 min	Google
ExSc5	Sc5:AiCEF	N/A	YES	20 min	–

Table 13 Demographics of the experts

Countries	#
Greece	2
Austria	1
Italy	1
Belgium	2
Poland	1
Spain	1
Romania	1
Portugal	1
Czechia	1
Netherlands	2
France	2
Finland	1

(a) Countries of origin of the experts

Sector	#
Governmental	6
Energy	2
ICT	4
Critical Infra	1
aw enforcement	1
Education	1
Other	1

(b) Sector that experts are working in

Seniority	#
Novice (small exercises)	5
Medium (medium scale exercises for a few years)	3
Expert (EU level, cross country exercises)	4
Senior (large-sized exercises)	1
None	3

(c) Seniority self-assessment

1. How much time did you invest in the Scenario Content Development?
2. How do you define the scope/objectives of the exercise?
3. How do you define the scenario content?
4. What tools did you use to create the scenario or define the objectives if any?

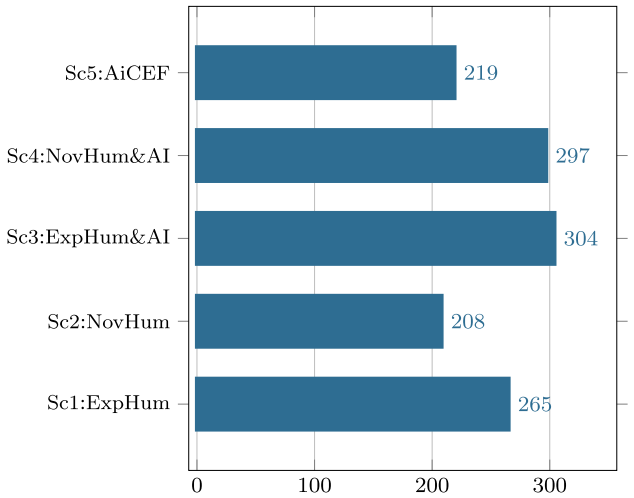


Fig. 11 Overall performance of evaluated scenarios based on total score

Finally, evaluators were asked to rank AI-powered tools as follows:

- Rank the following AI-powered tools that could be created to support the design and implementation of future cyber exercises:
  - Automated extraction of Exercise Objects (Incidents, Injects) from unstructured information and DB storage
  - Lead generation for trend prediction of Training Topics
  - Automated Enrichment's of content to match realistic patterns and relationships of known Attackers
  - Automated Cyber Exercise Script/Scenario Generation

5.3 Results analysis

The analysis of the input provided a good understanding of the strengths and potential areas for improvement of AiCEF. It also provided better insight into the exercise Scenario creation process, with good inputs for future improvement based on the experience of real EPs (Fig. 11).

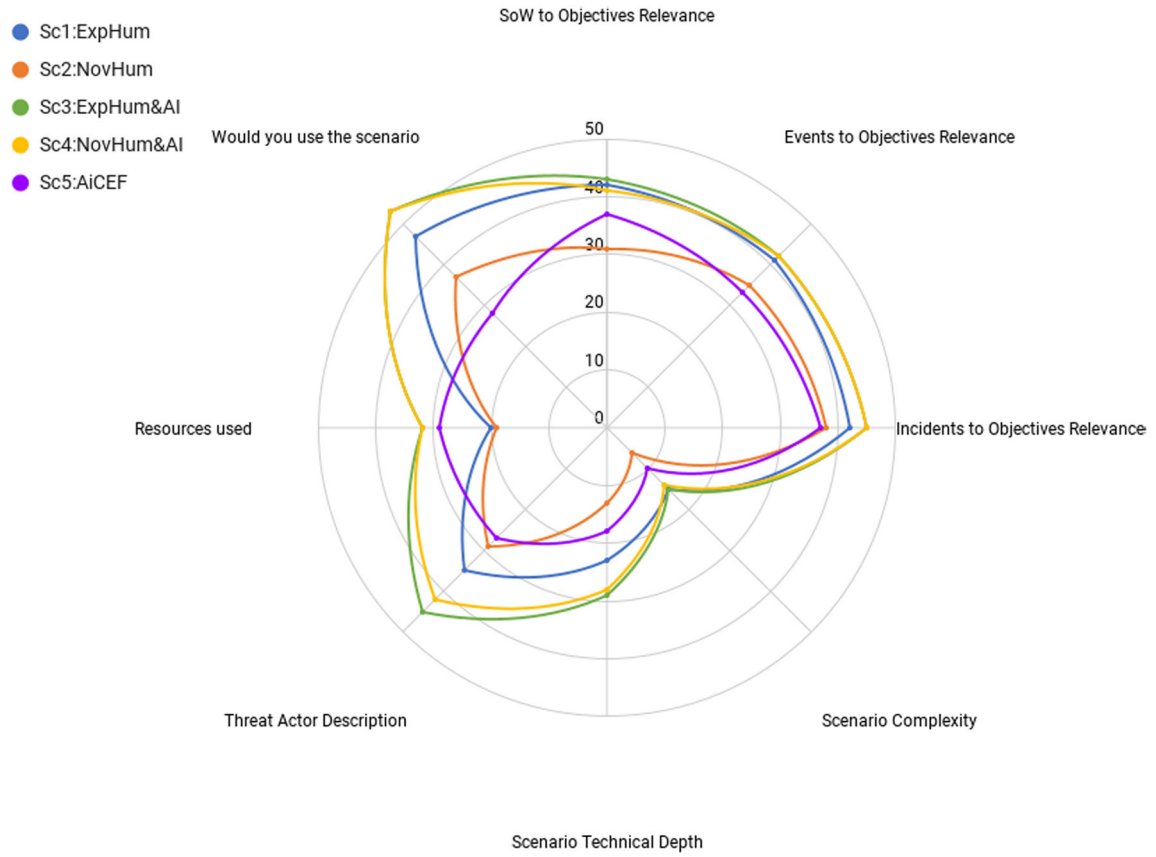


Fig. 12 Scenario evaluation parameters

Based on analysis of the provided input, we can safely conclude that both scenarios Sc3:ExpHum and AI and Sc4:NovHum and AI have scored higher than any other scenario with the help of AiCEF, see Fig. 14. Currently, the hybrid scenario generation approach of a human exercise planner using AiCEF outperforms a seasoned exercise planner, even when a planner is a novice. Furthermore, the Scripted Exercise Planner generated a relatively good Scenario (Sc5:AiCEF) that can be evaluated as equal, if not better, than that of a novice planner (SC2:NovHum) (Fig. 12).

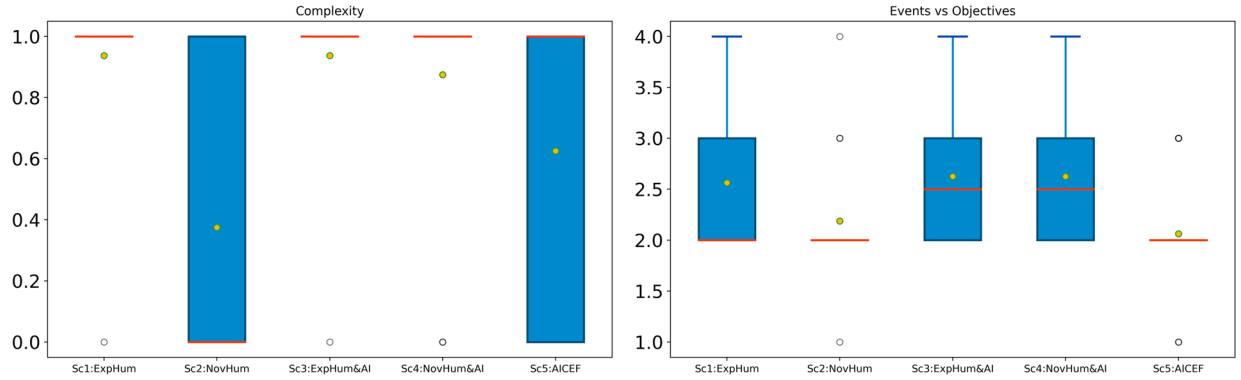
In what follows, we provide a breakdown of the parameters evaluated to highlight the strengths and weaknesses of using AiCEF based on the experts' input.

The use of AiCEF by a Scripted Exercise Planner performed well (top 3, outperforming humans) in *Relevant Resources*, *Events Relevance*, and *Scenario Technical Depth*. On the other hand, AiCEF did not perform as well in the following aspects: *Threat Actor Description*, *Scenario Complexity*, and *Incidents to Objectives Relevance*. The above can be justified by the fact that the raw generated content can include conflicting information or content that might not match the high-level context requested. After human curation, the content can be easily improved to compete with a seasoned exercise planner. In fact, AiCEF used by humans helped them excel in *Scenario Creation*, dominating all categories versus their human counterparts. The human expert using AiCEF (Sc3:ExpHum&AI) managed to create a better scenario 33,33% faster than his expert peer using regular tools (Sc1:ExpHum) (Fig. 13).

Nevertheless, the most impressive finding was that novice planners using AiCEF (Sc4:NovHum&AI) outperform a seasoned exercise planner (Sc1:ExpHum), as seen in Fig. 12, providing a good indication of the capabilities of the proposed framework. Note that the scenario performance developed by the novice planner with the help of AiCEF matches, among others, that of a Seasoned Planner in the question: *Would you use the scenario?*". Even more, evaluators could not distinguish the pure AI-generated content (ExSC5) based on Table 14, categorising the scenario as either hybrid or human-made. Indeed, the results were like those of a novice human planner.

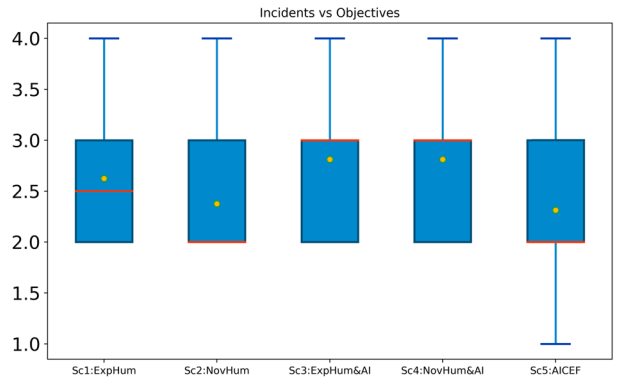
On the question: "*How do you define the scope/objectives of the exercise?"* most evaluators replied with two or more of the following options, with known incidents and lessons learnt along with risk assessment as the most prevalent replies (Fig. fig:score).

On the question: "*How do you define the scenario content?*" most evaluators replied with two or more options, with

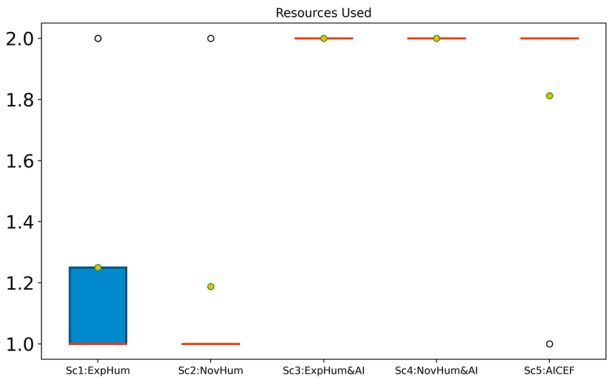


(a) Complexity

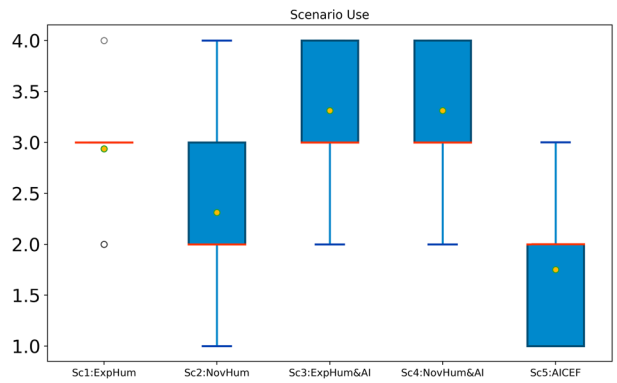
(b) Events vs Objectives



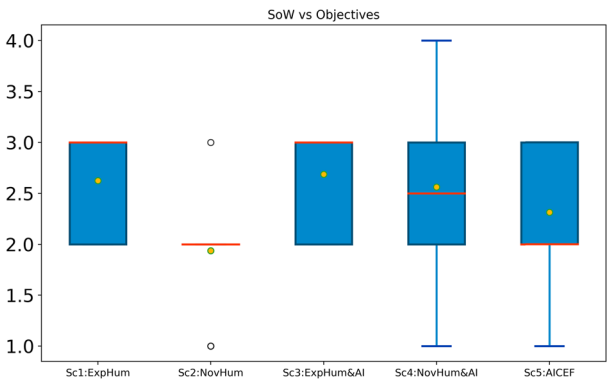
(c) Incidents vs Objectives



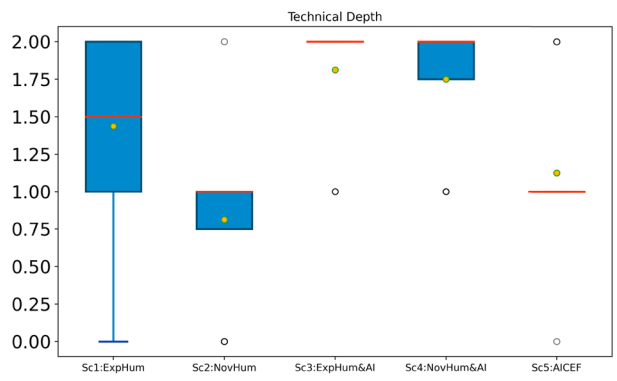
(d) Resources used



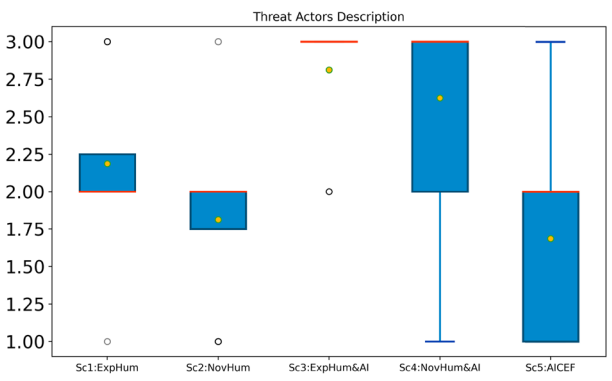
(e) Scenario use.



(f) SoW vs Objectives



(g) Technical depth



(h) Threat actors' description

Fig. 13 Score range for the Q1–8 of 16 evaluators

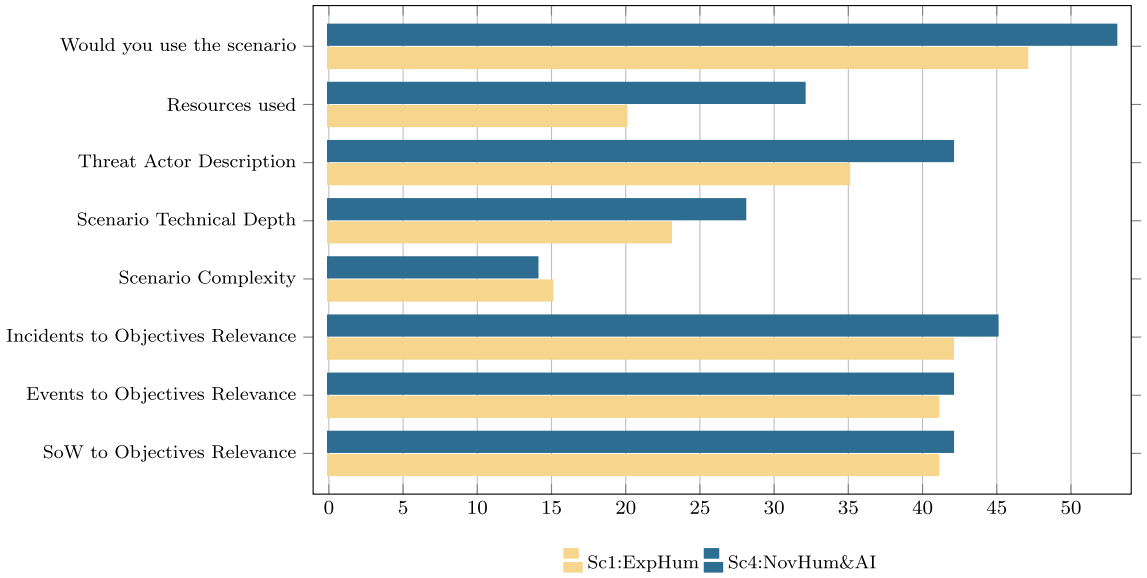


Fig. 14 Novice planner with AiCEF (Sc4:NovHum&AI) versus senior exercise planner (Sc1:ExpHum)

Table 14 Turing test to evaluate the performance of AI

Scenario	Human	AI and human	AI
Sc1:ExpHum	1	13	2
Sc2:NovHum	9	2	5
Sc3:ExpHum&AI	0	9	7
Sc4:NovHum&AI	1	10	5
Sc5:AiCEF	6	5	5

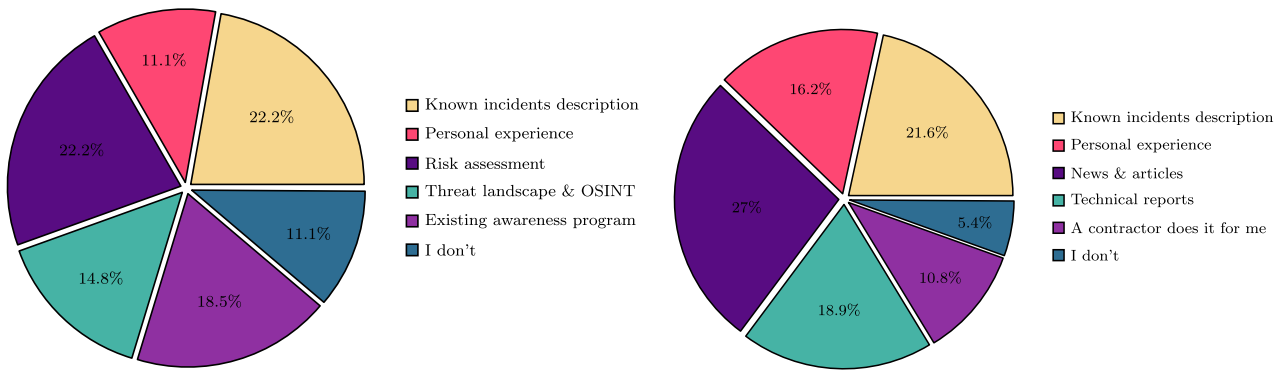
news and articles being the most important source followed by the known incident option (Fig. 15).

The evaluators replied to the question "*How much time do you invest in the Scenario Content Development?*" with an average of 53 h. This means that tools which can improve the CSE scenario content development process by reducing time without compromising the quality could be of great use.

Finally, for the question "*What tools did you use to create the scenario or define the objectives if any?*", the responses varied between Google Search, Cyber Security (News) websites, MS Office, and Internet/Table Top Research.

6 Conclusions and future work

The shortage of cybersecurity experts and awareness is a well-known and big worldwide challenge. CSEs can address some of the aspects of this problem; however, the shortage of experts to develop new CSEs coupled with the timeliness and relevance of the developed CSEs requires novel solutions. In this work, we try to fill in this gap by facilitating the



(a) Experts' responses to "How do you define the scope/objectives of the exercise?" question.

(b) Experts' responses to "How do you define the scenario content?" question.

Fig. 15 Experts' responses

work of EPs with the use of AI. To this end, we developed a novel AI-powered exercise generation framework called AiCEF, which generates structured exercise scenarios that reflect the current or future threat level that an organisation faces, including potential threat actors and TTPs. Moreover, it generates scripted events that could happen in the context of a real attack against a specific organisation belonging to one of the NIS2 critical infrastructure sectors. AiCEF also identifies and describes artefacts that could accompany the exercise scenarios. To this end, AiCEF uses a new ontology that we built, named CESO, and with which we were able to generate structured exercise scenarios that can be both machine and human-readable.

Our proposed methodology and developed tools can provide tangible qualitative and quantitative added value in CSE development and Cyber Awareness in various ways. For instance, in our experiment, the total time for the CSE scenario generation is decreased by 33.33% without impacting the quality. In fact, AiCEF improves the quality of CSE scenario generation for an inexperienced/novice EP by elevating the generated scenario quality to the same level as an experienced EP. Finally, the relevance of proposed CSE scenarios is aligned with that of the current threat landscape, as indicated by evaluating all the generated scenarios using AiCEF.

While AiCEF might be rather efficient, there is room for various improvements. For instance, for operational usage, more sources must be parsed (ex., threat reports and alerts) to generate more diverse scenarios. While Generative Pre-trained Transformers (GTP-2 and GTP-3) [4] [5] might create a textual output of very good quality, it would be even better if the text synthesiser were based only on Cyber Security related resources so that the generated text is even more relevant and uses, e.g. better technical terms. As indicated in the evaluation, AiCEF could be benefited from further improvements to enhance the threat actor description section. Finally, we plan to enhance AiCEF to detect the Cyber Kill Chain phases automatically using NER and create relevant CSE injects for a number of popular categories like phishing while also automating the inject description and content generation using AI-powered text synthesis.

Acknowledgements This work was supported by the European Commission under the Horizon 2020 Programme (H2020), as part of the project CyberSec4Europe (https://www.cybersec4europe.eu) (Grant Agreement no. 830929), and Horizon Europe Programme, as part of the project LAZARUS (https://lazarus-he.eu/) (Grant Agreement no. 101070303). The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

Funding Open access funding provided by HEAL-Link Greece.

Data availability The used data are publicly available.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval The authors declare full compliance with ethical standards. This article does not contain any studies involving humans or animals performed by any of the authors.

References

1. Adams, W.J., Gavas, E., Lacey, T.H., Leblanc, S.P.: Collective views of the NSA/CSS cyber defense exercise on curricula and learning objectives. In: CSET (2009)
2. Zacharis, A., Gavrila, C.P.R.: AI-assisted cyber crisis management exercise content generation: Modelling a cyber conflict. In: 15th International Conference on Cyber Conflict (CyCon 2023). IEEE (2023)
3. Augustine, T., Dodge, R.C., et al.: Cyber defense exercise: meeting learning objectives thru competition. In: Proceedings of the 10th Colloquium for Information Systems Security Education (2006)
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, S., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020a)
5. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, S., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020b)
6. Conklin, A.: The use of a collegiate cyber defense competition in information security education. In: Proceedings of the 2nd Annual Conference on Information Security Curriculum Development, pp. 16–18 (2005)
7. Conklin, A.: Cyber defense competitions and information security education: An active learning solution for a capstone course. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), vol. 9, pp. 220b–220b. IEEE (2006)
8. Craig, R.T.: Generalization of Scott's index of intercoder agreement. Publ. Opin. Q. **45**(2), 260–264 (1981). https://doi.org/10.1086/268657
9. Dewar, R.S.: Cybersecurity and Cyberdefense Exercises. Tech. rep, ETH Zurich (2018)
10. Dodge, R., Ragsdale, D.J.: Organized cyber defense competitions. In: Proceedings of IEEE International Conference on Advanced Learning Technologies, pp. 768–770. IEEE (2004)

11. Dodge, R., Hay, B., Nance, K.: Standards-based cyber exercises. In: 2009 International Conference on Availability, Reliability and Security, pp. 738–743. IEEE (2009)
12. European Commission: Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive) (Text with EEA relevance). https://eur-lex.europa.eu/eli/dir/2022/2555 (2022)
13. Furtună, A., Patriciu, V.V., Bica, I.: A structured approach for implementing cyber security exercises. In: 2010 8th International Conference on Communications, pp. 415–418. IEEE (2010)
14. Granåsen, M., Andersson, D.: Measuring team effectiveness in cyber-defense exercises: a cross-disciplinary case study. Cognit. Technol. Work **18**(1), 121–143 (2016)
15. Green, A., Zafar, H.: Addressing emerging information security personnel needs. a look at competitions in academia: Do cyber defense competitions work. In: AMCIS 2013 Proceedings, vol. 1, p. 257 (2013)
16. Gurnani, R., Pandey, K., Rai, S.K.: A scalable model for implementing cyber security exercises. In: 2014 International Conference on Computing for Sustainable Global Development (INDIACom), pp. 680–684. IEEE (2014)
17. of Homeland Security UD: DHS Cyber TTX for the healthcare industry. https://www.hsdl.org/?abstract&did=789781 (2013)
18. ISO Central Secretary: Societal security - guidelines for exercises. Standard ISO22398:2013, International Organization for Standardization, Geneva, CH. https://www.iso.org/standard/50294.html (2013)
19. Karagiannis, S., Magkos, E.: Engaging students in basic cybersecurity concepts using digital game-based learning: computer games as virtual learning environments. In: Advances in Core Computer Science-Based Technologies, pp 55–81. Springer (2021)
20. Karjalainen, M., Kokkonen, T., Puuska, S.: Pedagogical aspects of cyber security exercises. In: 2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), pp. 103–108. IEEE (2019)
21. Kick, J.: Cyber exercise playbook. Tech. rep, MITRE CORP BEDFORD MA (2014)
22. Li, Y., Liljenstam, M., Liu, J.: Real-time security exercises on a realistic interdomain routing experiment platform. In: 2009 ACM/IEEE/SCS 23rd Workshop on Principles of Advanced and Distributed Simulation, pp. 54–63. IEEE (2009)
23. Liljenstam, M., Liu, J., Nicol, D.M., Yuan, Y., Yan, G., Grier, C.: Rinse: the real-time immersive network simulation environment for network security exercises (extended version). Simulation **82**(1), 43–59 (2006)
24. Lockheed Martin: The Cyber Kill Chain. https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html (2011)
25. MacIntyre, R.: Penn treebank tokenizer (sed script source code) (1995)
26. Mattson, J.A.: Cyber defense exercise: A service provider model. In: IFIP World Conference on Information Security Education, pp. 81–86. Springer (2007)
27. Mink, M., Freiling, F.C.: Is attack better than defense? teaching information security the right way. In: Proceedings of the 3rd annual conference on Information security curriculum development, pp. 44–48 (2006)
28. MITRE: CVE. https://cve.mitre.org/ (1999)
29. MITRE: MITRE ATT&CK. https://attack.mitre.org/ (2022)
30. Mullins, B.E., Lacey, T.H., Mills, R.F., Trechter, J.E., Bass, S.D.: How the cyber defense exercise shaped an information-assurance curriculum. IEEE Secur. Privacy **5**(5), 40–49 (2007)
31. Mullins, B.E., Lacey, T.H., Mills, R.F., Trechter, J.M., Bass, S.D.: The impact of the nsa cyber defense exercise on the curriculum at the air force institute of technology. In: 2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07), pp. 271b–271b. IEEE (2007b)
32. OASIS OPEN: STIX version 2.1. https://www.oasis-open.org/standard/stix-version-2-1/ (2021)
33. Pastuszuk, J., Burek, P., Ksieopolski, B.: Cybersecurity ontology for dynamic analysis of it systems. Procedia Comput. Sci. **192**, 1011–1020 (2021)
34. Patriciu, V.V., Furtuna, A.C.: Guide for designing cyber security exercises. In: Proceedings of the 8th WSEAS International Conference on E-Activities and information security and privacy, World Scientific and Engineering Academy and Society (WSEAS), pp. 172–177 (2009)
35. Planning, M.E.: Directors's Guideline for Civil Defence Emergency Management Groups, wyd. Ministry of Civil Defence & Emergency Management, Wellington (2008)
36. Rursch, J.A., Luse, A., Jacobson, D.: It-adventures: A program to spark it interest in high school students using inquiry-based learning with cyber defense, game design, and robotics. IEEE Trans. Educ. **53**(1), 71–79 (2009)
37. Samejima, M., Yajima, H.: It risk management framework for business continuity by change analysis of information system. In: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1670–1674. IEEE (2012)
38. Sangster, B., O'Connor, T., Cook, T., Fanelli, R., Dean, E., Morrell, C., Conti, G.J.: Toward instrumenting network warfare competitions to generate labeled datasets. In: CSET (2009)
39. Scarfone, K.A., Grance, T., Masone, K.: Sp 800-61 rev. 1. computer security incident handling guide (2008)
40. Schepens, W., Ragsdale, D., Surdu, J.R., Schafer, J., Port, R.N.: The cyber defense exercise: an evaluation of the effectiveness of information assurance education. J. Inf. Secur. **1**(2), 1–14 (2002)
41. Schepens, W.J., James, J.R.: Architecture of a cyber defense competition. In: SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483), vol. 5, pp. 4300–4305. IEEE (2003)
42. Schweitzer, D., Gibson, D., Collins, M.: Active learning in the security classroom. In: 2009 42nd Hawaii International Conference on System Sciences, pp. 1–8. IEEE (2009)
43. Sommestad, T., Hallberg, J.: Cyber security exercises and competitions as a platform for cyber security experiments. In: Nordic conference on secure IT systems, pp. 47–60. Springer (2012)
44. Tobey, D.H.: A vignette-based method for improving cybersecurity talent management through cyber defense competition design. In: Proceedings of the 2015 ACM SIGMIS Conference on Computers and People Research, pp. 31–39 (2015)
45. Tsinganos, N., Mavridis, I.: Building and evaluating an annotated corpus for automated recognition of chat-based social engineering attacks. Appl. Sci. **11**(22), 10871 (2021)
46. Vigna, G.: Teaching network security through live exercises. In: IFIP World Conference on Information Security Education, pp. 3–18. Springer (2003)
47. Wen, S.F., Yamin, M.M., Katt, B.: Ontology-based scenario modeling for cyber security exercise. In: 2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), pp. 249–258. IEEE (2021)
48. White, G.B., Dietrich, G., Goles, T.: Cyber security exercises: testing an organization's ability to prevent, detect, and respond to cyber security events. In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences (2004), p. 10. IEEE (2004)

49. White, G.B., Williams, D., Harrison, K.: The cyberpatriot national high school cyber defense competition. IEEE Secur. Privacy **8**(5), 59–61 (2010)
50. Wilhelmson, N., Svensson, T.: Handbook for planning, running and evaluating information technology and cyber security exercises. Försvarshögskolan (FHS) (2011)