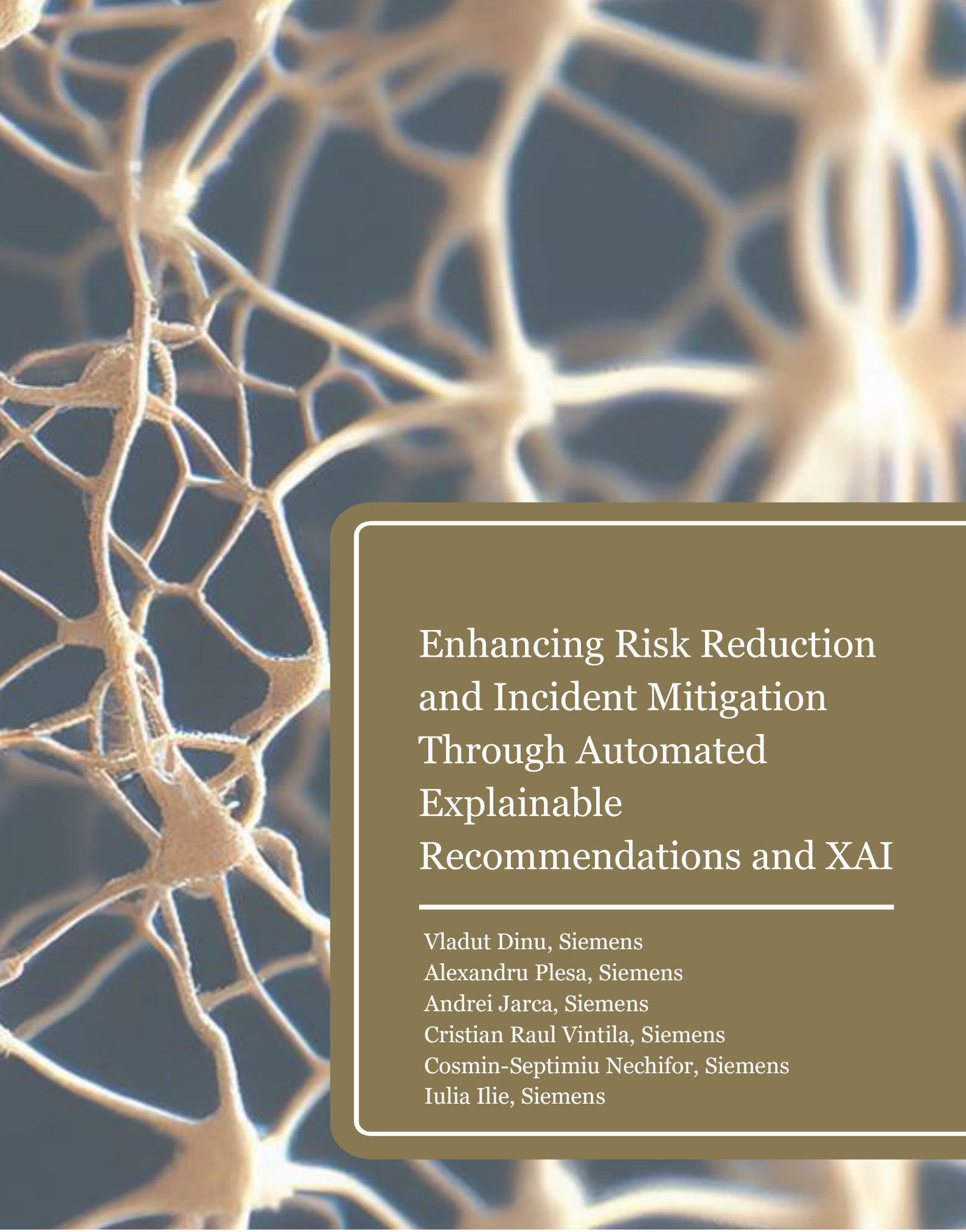

Enhancing Risk Reduction and Incident Mitigation Through Automated Explainable Recommendations and XAI

Vladut Dinu, Siemens
Alexandru Plesa, Siemens
Andrei Jarca, Siemens
Cristian Raul Vintila, Siemens
Cosmin-Septimiu Nechifor, Siemens
Iulia Ilie, Siemens

Co-funded by the European Union

This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under the Grant Agreement No. 101073909.

Enhancing Risk Reduction and Incident Mitigation Through Automated Explainable Recommendations and XAI

Vladut Dinu, Alexandru Plesa, Andrei Jarca, Cristian Raul Vintila, Cosmin-Septimiu Nechifor, and Iulia Ilie (Siemens)

We propose a framework of Risk Reduction and Incident Mitigation with integrated Explainable Artificial Intelligence (XAI) techniques to be used in post-crisis optimization of decision making for mitigating cyber-physical risks of the critical infrastructure. This approach can be used for enhancing transparency in threat detection, providing human-in-the-loop with clear insights into decision-making processes.

1. Introduction

Artificial Intelligence (AI) has rapidly evolved over the last two decades and it is continuously integrated into more use cases. The challenge is to create trust and reputation towards specific AI tools in the context of high-stakes decision making in safety-critical domains. To this end, Masakowski et al., 2022 [1] and Velev et al., 2023 [2] reinforce the importance of decision-makers and for them to understand the mechanisms behind AI frameworks. Moreover, they should be able to conceptualize a human-machine connection and interpret the reason behind an AI prediction.

Focusing on cyber-physical crisis, as an input, the Risk Reduction and Incident Mitigation framework of ATLANTIS offers recommendations on optimized sequences of actionable countermeasures for recovery. The human-in-the-loop is responsible to work towards an acceptable result, using meaningful insights into the decision-making process.

This paper introduces an approach for assessing the resulting recommendations with appropriate explanations enabling the humans-in-the-loop to accept or steer the result into the desired direction, supporting a continues learning system. The integration of XAI techniques is combined with Large Language Models (LLMs) [9] into the evaluation of the recommendation process to understand the multi-dimensional nuances of the explanations and facilitate critical thinking due to the detailed breakdown of the information.

2. Recommender Systems in Cybersecurity

The field of recommender systems in security, particularly in the context of risk reduction and incident mitigation, will increasingly be focused on the integration of automated and explainable recommendations through Explainable Artificial Intelligence (XAI). This evolution addresses the critical need for transparency and understandability in security-related decision-making processes. Recommender systems need to be context aware and have access to domain knowledge to present results in conformance to operational policies

and security strategies, as highlighted by Yang Li et al, 2021 [3]. Current models often are very specialized, limiting their effectiveness in the extended context of systems-of-systems. This situation highlights the need for these systems to evolve to a general purpose according to domain-specific knowledge, optimized by the domain specific strategies and regulated by domain-specific policies. Additionally, the need for better-informed decision-making by human operators calls for automated methods to enhance operational transparency and trust.

3. The Role of Explainable Recommendations

Risk Reduction and Incident Mitigation is a post crisis mechanism that suggests short-medium actions as a response to emerging situations that have been assessed in the context of cyber-physical safety and security of post crisis mechanism. The results are actionable, effective, efficient and context aware recommendations for risk reduction and incident mitigation.

The termination of the crisis makes it possible to collect a complete dataset of comprehensive information to understand the enhanced situational picture during the complete timespan of the crisis. However, we need to take the possible challenges that may appear into consideration, regarding the quality and availability of the data in the context of critical scenarios [2][4]. Since training data may be tampered or inadequate, an LLM/Multimodal LLM (MLLM) is capable to decipher the artefacts and thus, bypass this impending concern. Additionally, by levereging the Hypervision Tool and Knowledge Base framework, we are able to address any issues that may appear with privacy and security, due to the Big Data ecosystem [2][5][6].

The Hypervision tool is based on the Crimson product. It's a distributed solution composed by a central server which store the situation and manage users rights access, by a desktop application which allows to have a Common Operational Picture (COP) of the current situation and by a mobile application for on-field users.

The knowledge Base serves as a centralized repository for information and data that can be used to provide support, guidance, and decision-making assistance. For instance, it contains the rules governing conformance criteria, which are essential for the conformance checking process.

Relevant threats, vulnerabilities and incidents that critical infrastructure (CI) assets have been exposed to, are analysed and processed to effective recommendations of actionable short-to-medium term strategies. Interdependence of cyber-physical and human systems, the way how cyber threats impact physical operations and vice versa, are of great importance when making sense of how attacks have affected the critical infrastructure assets. The modelling or structured description of the critical infrastructure interdependencies defines the context of the ecosystem where an attack has been presumably executed by highly skilled motivated, well-resourced individuals or groups.

Managing information from the variety of different types of infrastructure, as outlined in ATLANTIS, as being systems of systems that are cross domain and cross border, operated by a high number of decentralized entities, is providing a more general domain-specific context. Incorporating domain-specific explicit information directly into models as

transparent as possible is realizable using tools for fine-tuning leading for enhancing model's performance on domain-specific tasks. The integration of XAI techniques is combined with LLM into the evaluation of the recommendation process to understand the multi-dimensional nuances of the explanations and facilitate critical thinking of humans-in-the-loop due to the detailed breakdown of the information.

4. The Research and Development Path in ATLANTIS

ATLANTIS embedded a powerful architecture to process and analyse digital information towards implementing cyber-physical protection and defence strategies. The Risk Reduction and Incident Mitigation framework combines the efficacy of recommendation engines with the ML NLP processing using BERT [7] Models, classifiers, embeddings, LLM and LIME to explain the output.

The proposed architecture of the Risk Reduction and Incident Mitigation Framework (RRIM; see Figure 1) is designed to interconnect among the various tools to mitigate incidents that are the Decision Support System (DSS), the Hypervision Tool, Tools to Fight Disinformation, the Human in Vicinity (HiViC), and the Knowledge Base. The interplay of this tools with the RRIM is iterative and enhances post-crisis decision-making capabilities due to risk mitigation strategies and experiential learning. The integration of human oversight introduces the collaborative work of computational intelligence and human judgement of data-driven and intuitively guided decision-making.

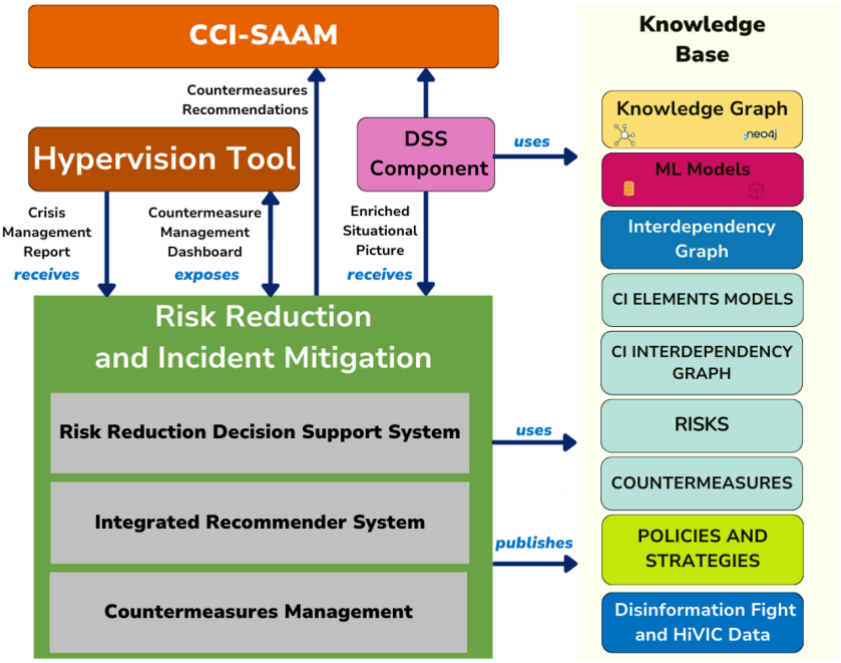

Figure 1. ATLANTIS RRIM architecture.

The information from the analysed crisis is ingested by the Data Input Layer from the various sources and is pre-processed to consolidate a robust dataset. The holistic reports of the Hypervision Tool are processed iteratively. A mechanism of data curation and data

imputation is put in place to synergistically use the Enriched Situational Picture in STIX [12] format, time series, text, images, and other data from the complementary sources.

The Data Input Layer is part of the Risk Reduction Decision Support System where the consolidated dataset is processed to manage the risks and other relevant features available from data describing the crisis in the declared time window. The Risk Assessment Layer uses AI methods to identify and evaluate potential risks. The Decision Support Layer interprets semantically enhanced data from the historical knowledge to provide comprehensive meta-information, enabling a thorough understanding of the security situation. It uses all pertinent context with respect to the knowledge of past attacks and interdependencies, to form an accurate picture of the current security status and potential influencing factors.

The list of attack descriptions during crisis known from the presented sources are evaluated against expert knowledge from the Knowledge Base to identify countermeasures. Candidates qualify by their potential effectiveness against the identified risks, considering the various factors of the analysed situation. Candidates who qualify are analysed by the Integrated Recommender Engine together with the meta-information that reflects the priorities and effectiveness in similar context.

The Integrated Recommender System leverages machine learning algorithms to identify overlooked risks, suggests optimal decisions based on similar past scenarios and offer tailored countermeasures. The IRS enables a more comprehensive and adaptive approach to risk management and incident mitigation with identified countermeasures. The Risk Optimization Recommender System processed the information to suggest optimized approaches with respect to risk management and contribute to the mitigation strategies. Matching the Policies & Strategies from the Knowledge Base and Integrated Recommender System potential countermeasures are identified and labelled with human in the loop.

Each identified countermeasure is then evaluated and scored based on several factors with human in the loop. This could involve considering each countermeasure's potential, time requirement, and any other relevant criteria. The scoring process might also consider historical data on the performance of similar countermeasures in similar contexts.

The countermeasures are reassessed for optimization and prioritization with human in the loop. The optimization process might involve adjusting or combining countermeasures to improve their effectiveness or efficiency. The prioritization process then ranks the optimized countermeasures based on their scores and any other relevant criteria, such as the severity of the risk they address.

Prioritized list of optimized countermeasures is prepared as a set of recommendations with human in the loop. These recommendations are translated to actionable countermeasures and delivered to the relevant systems or stakeholders to guide their response to the identified risks or incidents. Humans-in-the-loop are supported in all steps by LLMs to create a better understanding of the data presented and to reduce the overwork that might occur [8].

A use-case can be presented as follows: the sequence of event during a crisis is processed in a dedicated pipeline. The situational picture can contain multiple types of data, from text to pictures, even videos. The system transforms input into vectors of values, called embeddings that can be easily used and stored for further processing. After this step is done, the risk reduction decision support system handles the data semantically and presents a list of relevant items to the Integrated Recommender System to recommend possible

countermeasures. The generated recommendations are presented in a web user interface for the human-in-the-loop system to select and rate the counter measures, based on its domain knowledge, expertise, and the explanation that the Explainable AI (XAI) module has ensembled.

The XAI Module contains algorithms of LIME [10] and SHAP [11] that can add explainability information about the choices of the involved AI Models. A Large Language Model (LLM) will aggregate the explainability information to generate human-friendly explanations. This LLM is fine-tuned on cybersecurity data to create quality output. Basically, the model is protected and continuously fine-tuned with the incoming and processed sequences of events to enhance its efficiency in this specific context.

Having a domain expert with the responsibility to approve the incoming counter measures for an event in a specific context, contributes to the improvement system of the recommender engine.

The human in the loop will rate the counter measures, as seen in Figure 2, using a rating system based on five stars, where one star specifies that the counter measure is not applicable at all, and five stars means that the counter measure is perfect to be applied in the current context to help solving the event.

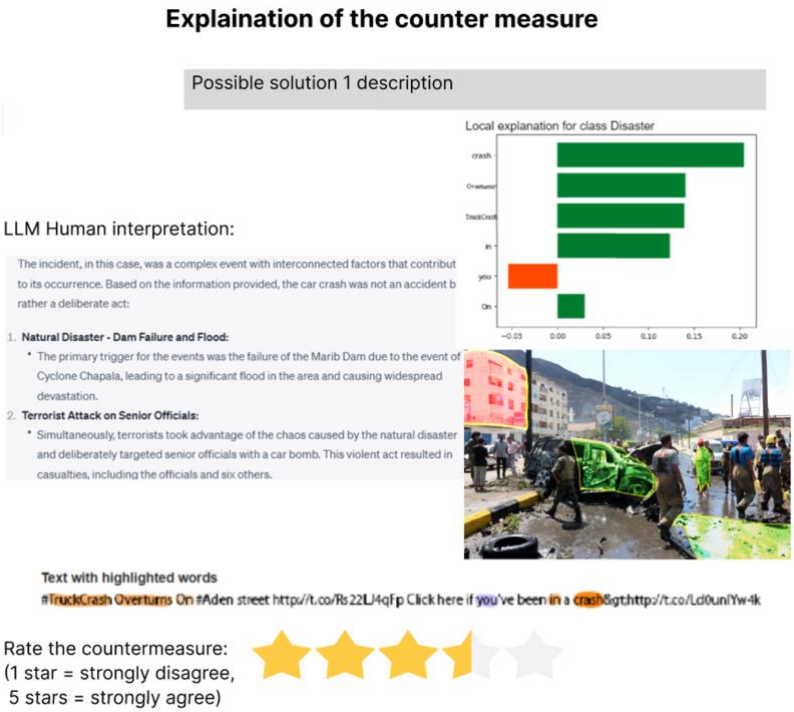

Figure 2. Web UI for Humans-in-the-Loop.

5. The Challenges and Barriers

One of the biggest challenges is to gather knowledge and correlated multi-modal datasets. Simply, there are multiple text datasets describing a cybersecurity event using human

language but there is no other type of data that can be correlated to create an enhanced situational picture of the event.

One example of a useful dataset is the Mitre ATT&CK database [13] which contains diverse topics of cybersecurity events, along with additional data such as countermeasures and mitigations solution, how to detect a specific event and other relevant links between these events. It also provides metadata such as specific platforms that can endangered by that event, the permission needed to do that attack and much more.

Having such detailed dataset with other types of data, such as images, video, and other types of data, can help developing the entire pipeline of recommendation and extract valuable information that can be further explained to determine which counter measure can be applied to solve an event in the specific context based on pre-recorded data.

6. The Benefits and Impact

There are plenty benefits to use the proposed approach. The domain expert exposed to this mechanism can evaluate the situation based on the explanations that the recommender system offers as an additional output and rate the possible counter measures. Using this approach, experts will continuously get involved into the background training system of the AI Models using a rating system. This was integrated into a human-in-the-loop feedback system for cyber-physical events, using a web user interface. Because this system is applied on real-world use-cases, it needs a domain expert to validate its possible counter measures to be further applied by the people that will act.

7. Future Outlook

The challenge of this system is whether to use or to integrate diverse types of data filtering methods, such as content-based, user-based filtering or machine learning models. These approaches are yet to be evaluated and determine how they are the best to be used. The first iteration of this system uses content-based filtering since it assures the best reusability in the implementation and can be used as is whenever it will be deployed.

Another useful and important feature would be that everything will be explained using human language and understanding. All the data gathered from the AI Models will be aggregated using a Large Language Model (LLM) that will output human-readable content based on all the processed information.

Due to the increasing amount of data that can be involved into this project, the solution proposed in this system is to use classifiers. The knowledgebase is already created based on the event types found in the Mitre ATT&CK dataset and the used BERT Models are specialised on security data. To assure an optimal response time for post-crisis events, the system will try to classify the event and its current context into one topic that the knowledgebase already is familiar with and, based on that topic, the recommender engine will extract specific information from that class of event, this way it will tighten the amount if information that needs to be circulated between the processing modules. Moreover, the

embeddings are stored in a separated Vector Database to skip the re-calculation part every time the pipeline will run.

The models involved into the backend of the pipeline will be integrated into a continuous learning system. They will get familiar with the incoming events and will learn them and their counter measures in such a way that they will perform better every time an event occurs. Moreover, this learning system will also contain key-performance indicators (KPIs) that will evaluate the efficiency and usability of the models.

8. Conclusions

In conclusion, the proposed Risk Reduction and Incident Mitigation framework, integrated with eXplainable AI (XAI) techniques, presents a comprehensive approach to addressing cyber-physical risks in critical infrastructure. The need for context-aware, domain-specific recommendations, the focus on understanding interdependencies between cyber and physical systems enhances the contextual relevance of recommendations. The framework emphasizes the importance of transparency and human-in-the-loop decision-making in high-stakes scenarios. By leveraging AI and machine learning models, the framework aims to provide actionable recommendations for post-crisis optimization, enhancing the overall security and safety of critical infrastructure.

References

[1] Masakowski, Y. R., (Ed.), 2022. Artificial Intelligence and Global Security: Future Trends, Threats and Considerations, Emerald Publishing.

[2] Velev, Dimiter & Zlateva, Plamena. (2023). CHALLENGES OF ARTIFICIAL INTELLIGENCE APPLICATION FOR DISASTER RISK MANAGEMENT. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XLVIII-M-1-2023. 387-394. 10.5194/isprs-archives-XLVIII-M-1-2023-387-2023.

[3] Li, Yang & Liu, Kangbo & Satapathy, Ranjan & Wang, Suhang & Cambria, Erik. (2023). Recent Developments in Recommender Systems: A Survey.

[4] Özyer, T., Bakshi, S., Alhajj. R., (Eds.), 2019. Social Networks and Surveillance for Society, Springer.

[5] Douglass, R., et al., (Eds.), 2023. IoT for Defense and National Security, Wiley.

[6] Gerunov, A., 2023. Risk Analysis for the Digital Age, Springer.

[7] Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[8] Zefan Cai, Baobao Chang, Wenjuan Han 2023. Human-in-the-Loop through Chain-of-Thought, arXiv.

[9] Naveed, Humza & Khan, Asad & Qiu, Shi & Anwar, Saeed & Usman, Muhammad & Barnes, Nick & Mian, Ajmal. (2023). A Comprehensive Overview of Large Language Models.

[10] Ribeiro, Marco & Singh, Sameer & Guestrin, Carlos. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 97-101. 10.18653/v1/N16-3020.

[11] Lundberg, Scott & Lee, Su-In. (2017). A Unified Approach to Interpreting Model Predictions.

[12] Sadique, Farhan & Cheung, Sui & Vakilinia, Iman & Badsha, Shahriar & Sengupta, Shamik. (2018). Automated Structured Threat Information Expression (STIX) Document Generation with Privacy Preservation. 10.1109/UEMCON.2018.8796822.

[13] Xiong, Wenjun & Legrand, Emeline & Åberg, Oscar & Robert, Lagerström. (2022). Cyber security threat modelling based on the MITRE Enterprise ATT&CK Matrix. Software and Systems Modelling. 21. 10.1007/s10270-021-00898-7.

Front cover image by Garik Barseghyan via Pixabay.
https://pixabay.com/users/insspirito-1851261