An Overview of AI Techniques in DYNABIC Project

Ciprian Paduraru[1], Rares Cristea[1] and Alin Stefanescu[1,2]

[1]University of Bucharest, Romania
[2]Institute for Logic and Data Science, Romania

Abstract

We present an overview of the AI techniques used in Horizon Project No. 101070455 DYNABIC (Dynamic business continuity and response of critical systems against advanced cyber-physical threats). The project, which started in December 2022 and spans three years, aims to enhance European critical services' resilience and business continuity against cyber-physical threats. The project focuses on real-time prediction, assessment, and mitigation of threats, as well as automated optimization and orchestration strategies for response and prevention. Demonstrations are conducted in four industry-based case studies, simulating attacks on intelligent transportation services, electric vehicle charging stations, 5G telecommunication infrastructure, and healthcare services. Each infrastructure is modeled using a digital twin, which simulates business processes and allows user input for control. The paper focuses on one of the uses of AI within DYNABIC involving large-language models fine-tuned to assist cybersecurity experts.

Keywords

large language models, cyber-physical threats, resilience, digital twin, real-time threat mitigation,

1. Introduction

This paper gives an overview of the AI techniques used in the Horizon Project No. 101070455 DYNABIC[1] (Dynamic business continuity and response of critical systems against advanced cyber-physical threats). The main motivation is to share some of our findings from the project with other ongoing projects and to look for future cooperation and funding opportunities.

In short, the project started in December 2022 and lasts 3 years. The project aims to analyze, develop, and disseminate methods to increase European critical services' resilience and business continuity capabilities in the face of advanced cyber-physical threats. These objectives and motivations emerge from the reports showing how cyber-physical threats can lead to business disruption and cascading effects on interconnected critical infrastructures.

The final contributions of the project to this goal are: (a) predicting, assessing, and mitigating threats in real-time, and (b) responding to and preventing these attacks through automated optimization and orchestration strategies. The techniques are currently demonstrated in four industry-based case studies with different simulated attacks on the respective infrastructure:

- an intelligent transportation service
- an electric vehicle charging station
- a 5G telecommunication infrastructure
- a healthcare service.

Each of these infrastructures is modeled with a digital twin (DT), which simulates the business processes on the one hand, while acting as an actuator on the other, i.e., allowing input from the user of the DT to control these processes. The complete architecture is depicted in Figure 1.

The next section describes our research on developing a large-scale language-based modeling assistant to help humans detect, prevent, and mitigate cybersecurity attacks.

RuleML+RR'24: Companion Proceedings of the 8th International Joint Conference on Rules and Reasoning, September 16–22, 2024, Bucharest, Romania

ciprian.paduraru@fmi.unibuc.ro (C. Paduraru); cristea.rares96@gmail.com (R. Cristea); alin@fmi.unibuc.ro (A. Stefanescu)

[1]https://dynabic.eu

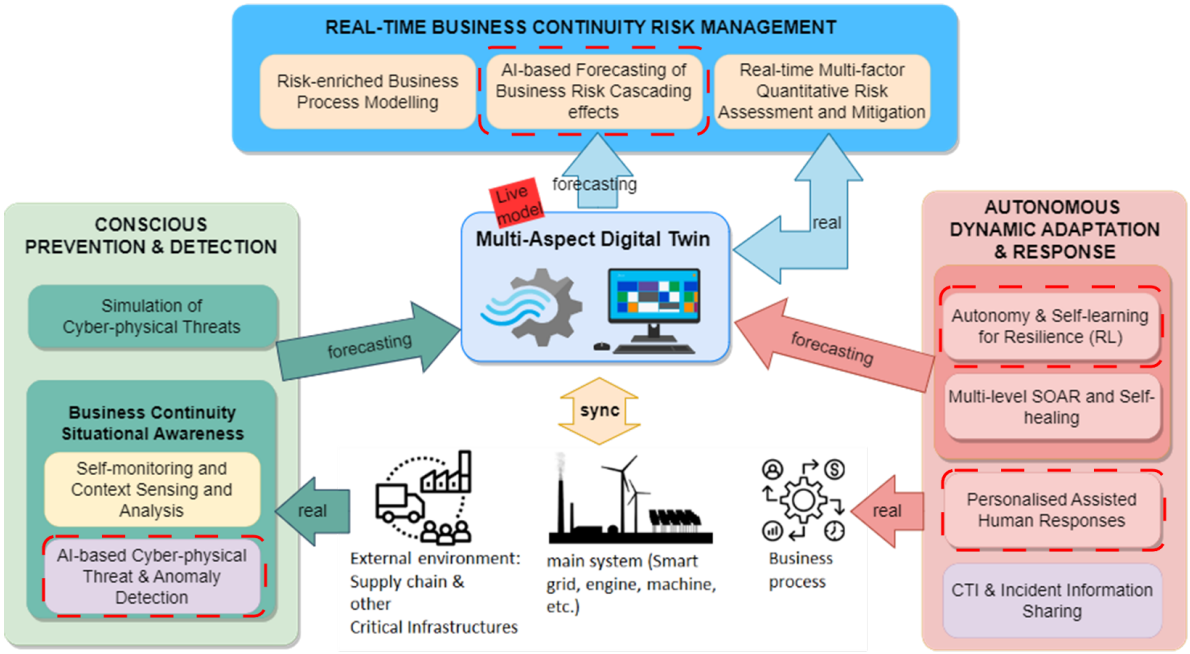

Figure 1: The architectural decomposition of the DYNABIC processes and components when used in a specific infrastructure (the architecture is taken from [1]). The AI components are marked with red rectangles.

2. Real-time personalized assistance for SoC operators

This subcomponent within the DYNABIC architecture focuses on providing real-time alerts and support to the Security Operations Center (SoC) specialists. The interaction takes the form of a live chatbot, where the human can collaborate with a personalized assistant that acts as an intermediate component between the human actions and the available systems and infrastructure for threat detection and prevention. We give a concrete usage example: First, the human SoC operator can be warned of a service interruption through automatic analysis of the logs (e.g., detecting many timeout responses on the user side. Then the human operator can ask questions in natural language on related topics such as database queries, graphs, and maps to finally find out that the cause is a DDoS (Distributed Denial of Services) attack. These tasks are solved by the assistant component by converting natural language questions into concrete actions, such as generating source code, executing, and retrieving data from the deployed systems it interacts with.

This work was disseminated in [2]. Technically, the core of the assistant support was built on a Large Language Model (LLM), namely Llama 3 8B [3] version. Our contribution was first to collect and clean a dataset of cybersecurity technical information consisting of about 5000 open research and book publications, and 3000 YouTube transcripts with courses from top universities, video blogs, and conference/workshop presentations. The base model was then fine-tuned to fit a cybersecurity domain and its new methods, attack types, and tool knowledge. The dataset[2] and model[3], including the source code to gather the dataset and fine-tune it, are publicly available. We started from Llama 3-Chat-7B base model and fine-tuned it using a dataset containing recent public cybersecurity content. The dataset, $\mathcal{D}$, was first exported as a series of JSON files. In addition, its metadata was stored on a MongoDB server. Finally, $\mathcal{D}$ was chunked and permanently stored in a vector database with FAISS (Facebook AI Similarity Search library from Meta).

To incorporate knowledge about the services and infrastructure with which the assistant can interact, we used the RAG (Retrieval Augmented Generation) technique. The security of operation and data processing was ensured by Llama Guard [4] and also by classical NLP techniques or basic network

[2]https://huggingface.co/datasets/unibuc-cs/CyberGuardianDataset
[3]https://huggingface.co/unibuc-cs/CyberGuardian

security methods such as ACL (Access Control List).

The DYNABIC team also evaluated the use of ChatGPT [5] via programmatic APIs in parallel, without fine-tuning. Both have advantages and disadvantages. For example, while ChatGPT-based assistants are easier to implement and prototype in an initial phase, they can raise security concerns when integrating knowledge about the infrastructure on which they operate.

Assistant personalizing and SSH (Social Sciences and Humanities). The SoC operators may have different expertise, communication preferences, or security risks when interacting with the deployed systems (either intentionally or due to lack of expertise or physical characteristics). Each of these components is analyzed and assessed using a questionnaire for each human operator. These are then integrated into the fine-tuned LLM-based assistant using prompt templates, ACLs, read/write permissions, and source code execution permissions. For example, a user who raises security concerns according to the outcome of the questionnaire will not be allowed by the assistant to execute source code in the infrastructure without the approval of their manager. The pipeline of evaluation, mining the user profile and assigning rights is shown in Figure 2.

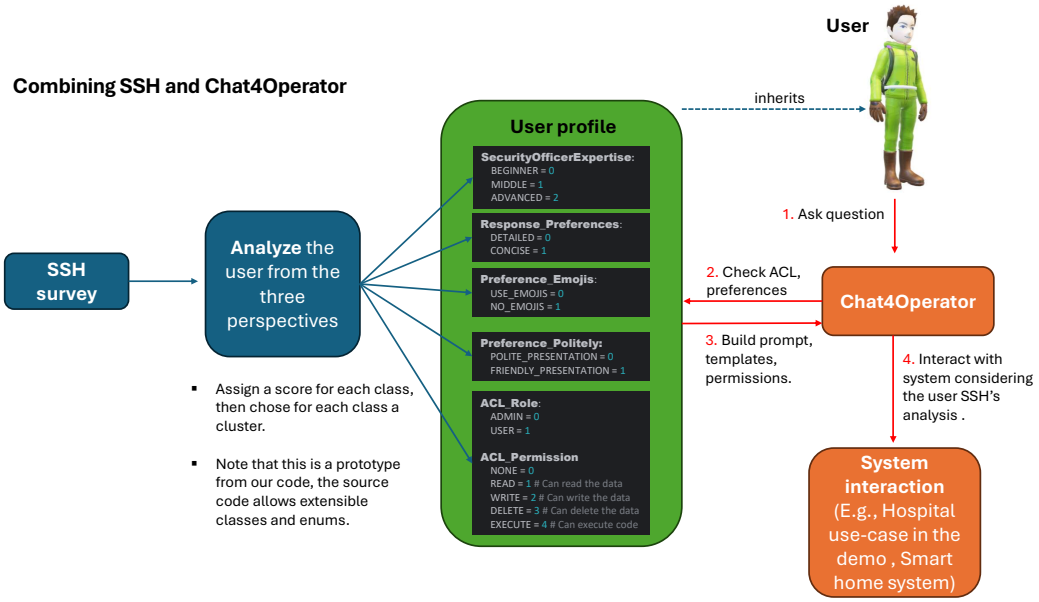

Figure 2: The profiling of an SoC operator begins with a survey in which the user is analyzed under various aspects such as security risks, expertise, etc. Based on this, the user profile is created and various rights (e.g., for code execution) are assigned. The operator uses these when interacting with the deployed system during the conversation with the fine-tuned chatbot assistant.

To cluster the human profiles and dynamically adapt the questionnaires at runtime based on their answers, the team has developed two AI-based solutions that are also publicly available: a) classification with a classical path-finding adapted algorithm that relates the distances to the profiles depending on the operator's answers [6] and b) a reinforcement learning (RL) algorithm that adds a background reward system to the previous solution [7].

The assistant is further personalized to include a head-like avatar with animation and voice (based on Unreal's Metahumans[4]), depending on the user's preferences so that the conversation becomes as human-like as possible.

[4]https://www.unrealengine.com/en-US/metahuman

Evaluation We sketch below the quantitative and qualitative evaluation methods used for this component, called CyberGuardian, using two research questions (see more details in [2]).

- **RQ1: How is the fine-tuned model able to understand the cybersecurity domain?**
  In this sense we used GPT-4 as a *judge* and evaluated 5 topics in cybersecurity needed by SoC specialists: (a) protection of systems from security risks and malware, (b) cryptography, (c) configuration of security protocols such as firewalls, (d) intrusion detection systems (IDS), (e) network security infrastructure (firewalls, VPNs, web proxy, IDS/IPS), (f) investigation of data breaches and data leaks. Then, 20 questions from each category were selected and evaluated. The percentage of cases in which each of the compared models was preferred to the CyberGuardianLLM using GPT-4 as judge: Llama 2 7B - 26%, Llama 2 13B 39%, Llama 2-70B - 58%.
- **RQ2: How well is the overall CyberGuardian system able to meet user needs?**
  The takeaway is that with reasonable computational requirements needs, the fine-tuned model can inherit cybersecurity-related general knowledge close to the much more expensive models. The BLEU metric [5] compared the following 4 models and we obtained these values: Llama 2 7B - 31.07, Llama 2 13 B - 41.97, Llama 2 70B - 71.42, and CyberGuardian - 65.91 (see more details in [2])

3. Risk mitigation, recommendation, and DYNABIC adaptive intelligence

To provide dynamic, autonomous strategic recommendations for countermeasures in case of an incident, RL methods are used. In particular, a Deep-Q-networks (DQN) algorithm is trained with experience replay to react in a simulated environment covering concrete use cases. The dataset used for training and evaluation is based on one of the project demonstrations, the electric vehicle charging station use case, where common incidents such as DDoS and false data injection pose a real threat. The reward function is modeled based on the protected infrastructure and its business objectives, such as availability, costs, and image. The agent can act on the environment by taking actions such as blocking IPs, using a honeypot server, resetting data, or doing nothing.

To build trust in the RL-based methods and provide debugging capabilities, the team introduced a method named *Chat4XAI* [8], which provides explanations of the decision-making process of the agent described above using an LLM-based assistant. At each step, the method explains the observation and decisions made by the RL agent. The natural language description of each step is made by leveraging OpenAI's ChatGPT API and prompt engineering.

It is also important to note that the abstract representation of the underlying systems in each use-case is made by classic methods used in AI, such as knowledge graphs, objects, and time-series databases.

Acknowledgments

This research was supported by the European Union's Horizon Europe research and innovation program under grant agreement no. 101070455, project DYNABIC.

References

[1] E. Rios, E. Iturbe, A. Rego, N. Ferry, J.-Y. Tigli, S. Lavirotte, G. Rocher, P. Nguyen, H. Song, R. Dautov, W. Mallouli, A. R. Cavalli, The DYNABIC approach to resilience of critical infrastructures, in: Proc. of ARES'23, ACM, 2023.

[2] C. Paduraru., C. Patilea., A. Stefanescu., CyberGuardian: An interactive assistant for cybersecurity specialists using large language models, in: Proc. of ICSOFT'24, SciTePress, 2024, pp. 442–449.

[5] https://huggingface.co/spaces/evaluate-metric/bleu

[3] A. Dubey, et al., The Llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.
[4] H. Inan, et al., Llama Guard: LLM-based input-output safeguard for human-AI conversations, 2023. URL: https://arxiv.org/abs/2312.06674. arXiv:2312.06674.
[5] A. Bahrini, et al., ChatGPT: Applications, opportunities, and threats, 2023. URL: https://arxiv.org/abs/2304.09103. arXiv:2304.09103.
[6] C. Paduraru., R. Cristea., A. Stefanescu, Adaptive questionnaire design using ai agents for people profiling, in: Proc. of ICAART'24, SciTePress, 2024, pp. 633–640.
[7] C. Paduraru., C. Patilea., A. Stefanescu., RLHR: A framework for driving dynamically adaptable questionnaires and profiling people using reinforcement learning, in: Proc. of ICSOFT'24, SciTePress, 2024, pp. 633–640.
[8] A. Metzger, J. Bartel, J. Laufer, An AI chatbot for explaining deep reinforcement learning decisions of service-oriented systems, in: Proc. of ICSOC'23, ACM, 2023, pp. 323–338.