Responses to Reviewer Comments for "Experimenting with neurosymbolic AI for defending against cyber attacks"

March 6, 2025

Firstly, we would like to express our gratitude to the reviewers and editors for their thorough and constructive feedback on the paper. We have done our outmost to address the comments and we believe this has significantly improved the quality of the paper. Below we include the reviews received in full and detail how we have addressed suggested changes in red.

Editor Comments

In addition to the reviewers' comments, I would also recommend adding a discussion about the fact that often datasets for cyber attacks detection are heavily unbalanced (as most of the time everything works well and only very few times an attack is registered). This might be added as an additional point in challenge 2 or as a different challenge all together.

- We agree that the imbalance is an issue. A discussion point has been added under challenge 2.

Finally, I would also recommend the authors to have a look at Prof. Cavallaro's (from UCL) work on how to handle the problem of concept drift in the cyber security domain and how his approaches can be improved using neurosymbolic techniques.

- We have included Cavallaro's *INSOMNIA: Towards Concept-Drift Robustness in Network Intrusion Detection* (as well as a tool called Telosion) as ways of adressing concept drift under the discussion related to challenge 1.

- We have also included Cavallaros work when discussing how knowledge about assets and threats can be used to address concept drift in a neurosymbolic fashion under Use case 1.

Review 1

The paper presents an insightful exploration of neurosymbolic AI (NeSy) as a hybrid approach to addressing challenges in cybersecurity, particularly in incident detection and response within Security Operations Centers (SOCs). By combining symbolic reasoning with neural network-based learning, the authors effectively argue for the relevance of NeSy in overcoming the limitations of standalone connectionist or symbolic systems.

The paper is significant in its focus on cybersecurity using NeSy, especially from the perspective of transparency, a challenge that connectionist systems alone cannot adequately address. The authors clearly articulate their points and present their efforts coherently, supported by initial case studies. The integration of domain knowledge into NeSy pipelines highlights their dedication to bridging theoretical advancements with real-world applications.

The main limitation of the paper lies in the lack of extensive experiments. However, this is not unique to this work; the application of NeSy in cybersecurity is still in its infancy, and further research will require time to mature.

- We agree that the experiments are not close to practical/operational use. However, the key message of this work is that there is unexplored potential of NeSy in this domain. We have also chosen to show the breadth of the potential by including several experiments/ use-cases. We have added two tables that shows this breath by indicating which use cases / challenges each experiment addresses, and also made this limitation clear in the introduction to the experiment section, where We also give a clearer description of the selection criteria and objectives of the experiments.

Overall, the paper is well-written, coherent, and concise. While the inclusion of more real-world case studies would strengthen its impact, it is acknowledged that this is a complex and challenging task.

Review 2

This paper is an extended version of a conference paper. This paper focuses on neurosymbolic approaches in the domain of cybersecurity. The authors identify a set of challenges, and propose a set of neurosymbolic use-cases. The notion of ”SOC” (security operation centre) plays an important role in this paper: ”A SOC consists of people, processes, and tools. One of the objective of SOC is to detect and respond to threats and attacks...”. In fact, the paper initially describes the typical use of AI in a SOC, and it identifies challenges in such a context. Use-cases are about applying NeSy in the context of a SOC. Experiments with neurosymbolic methods in cyber security challenges are included. Overall, studying neurosymbolic approaches to detect and react to cyber attacks is novel and of interest for the NeSy community.

The 9 challenges are well described, and summarizing them offers a useful description that can be helpful to the average reader. However, overall, they are pretty shallow descriptions, and they might be of limited interest to people in the field.

- We have deliberately tried to keep the challenges fairly abstract in order to be sufficiently broad as well as to make it understandable for both NeSy and security researchers. It is hard to find the correct balance and We have tried to expand the discussion a bit for the following challenges, as we felt that this could help the reader: 1, 2, 3, 6, 8.

Compared to the conference version, this paper includes more experimental activity, thus going beyond a major limitation of the original paper.
I found the experiments interesting, even if some of them sound very artificial. Given the lack of maturity of NeSy in this field, I still find them useful.

- We agree with this and have attempted to have several different experiments to show the broad potential use. A similar comment was made by the first reviewer, where we have detailed this further.

Considering LLM (in some challenges) is attracting, given their current popularity. Use some colors to highlight the syntax in fig 10 and 11, for better readability.

- We have added syntax highlighting for figures 8, 9 and 10.

Conclusions: check ref 83, there might be a typo (I see a question mark there).

- Typo in reference fixed.

The authors are expected to cite and mention earlier work in the context of defending to adversarial attacks by neurosymbolic approaches: Melacci, S., Ciravegna, G., Sotgiu, A., Demontis, A., Biggio, B., Gori, M., & Roli, F. (2021). Domain knowledge alleviates adversarial attacks in multi-label classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12), 9944-9959.

- Our focus is on the use of AI to defend against cyber attacsk and we therefore consider approaches focusing on AI security to be outside the scope of this paper. We have clarified this in the introduction and the conclusion, including a citation to this particular paper.

Review 3

This paper is an extension of a study previously presented at the NeSy conference. The additional content, however, is not entirely new, as it builds on some of the experiments introduced at the conference and incorporates a review of another recent paper authored by the same researchers (ref 101).

Overall, the paper presents an interesting cyber-security scenario for neuro-symbolic techniques. However, it falls short of convincingly demonstrating the suitability of these techniques for the chosen experiments. Let's organize my main concerns and comments with the ordering of the paper.

Abstract

The abstract is too short and it must include the main hypothesis and conclusions.

- We have rewritten the abstract to include both the hypothesis and conclusion so that the core message is more clearly conveyed there as well.

Section 2

It must include subsections for the different phases of MAPE-K, namely: 2.1. Monitor, 2.2. Analyze, etc.

- We have added subsections for the different phases of MAPE-K. Note however that the plan and execute phases are joined into one section as this part of neurosymbolic AI is the less mature (with no challenges/use-cases in the execute phase identifed)

The reference to ENISA only needs to be cited once.

- This has been fixed.

Challenge 3 his challenge does not include any reference. Are there any existing approaches related to this challenge that could be cited?

- We have added two existing approaches to improve intrusion detection with knowledge in the discussion of Challenge 3:

  - Alice Bizzarri, Brian Jalaian, Fabrizio Riguzzi, and Nathaniel D Bastian. *A neuro-symbolic artificial intelligence network intrusion detection system.* In 2024 33rd International Conference on Computer Communications and Networks (ICCCN), pages 1–9. IEEE, 2024.
  - Anna Himmelhuber, Dominik Dold, Stephan Grimm, Sonia Zillner, and Thomas Runkler. *Detection, explanation and filtering of cyber attacks combining symbolic and sub-symbolic methods.* In 2022 IEEE Symposium Series on Computational Intelligence (SSCI), pages 381–388. IEEE, 2022.

Include a comment about the maintenance of the ontologies, especially SEPSES, which appears to be inactive or stuck as of 2023. In Plan&Execute there should be some comment about recommendation systems in cyber-security. This topic is addressed in section 4.5, so a connection should be made here.

- We have include a small discussion on recommender systems in the chapter for Plan&Execute when introducing challenge 8. We chose not to make a connection to the recommender system referenced in 4.5, as that system was used for analysis, while this section is focused on planning and executing a response action.

Section 3

Authors must include a paragraph or table explaining the different types of neuro-symbolic AI. Specifically, describe the different ways large language models (LLMs) and symbolic approaches can be combined to solve problems. The introduction to neuro-symbolic (nesy) approaches should be more formal, clearly describing the techniques that fall into each category and how they can be combined.

- We have added a new section *Neurosymbolic AI to defend against cyber attacks* to introduce the different types of neurosymbolic AI, and neurosymbolic AI as a whole. We moved the definition of the techniques from section 3 into the new section, while keeping the information on application in section 3. The content is structured using Kautz Taxonomy. We have also included more work on the combination of LLMs and NeSy (still using Kautz taxonomy).

It is recommended to include a final table relating the challenges to the use cases.

- We have added a new table (Table 1) relating use cases with challenges addressed. We have also mapped the experiments to this table
- Additionally, we added table 2 in the conclusion chapter as a more general overview relating use cases to challenges, experiments and the different phases of MAPE-K.

Similar to Section 2, this section must include subsections for the different phases of MAPE-K, namely: 3.1. Monitor, 3.2. Analyze, etc.

- We have subsections for the different phases of MAPE-K mirroring section 2.

Section 4

Authors must justify the selection of the experiments. Why have these specific use cases been chosen over others?

- We have rewritten the introduction to chapter 5 to give clearer description of the selection criterias and objectives of the experiments.

Section 4.1

The results of Figure 2 should be presented as a table instead of a figure.

- We changed figure 2 into a table (Table 2)

Does the training data for the baseline neural network (NN) include the NWS vs. IT feature? If not, then the results are not fair, and this should be addressed.

- The baseline network does not contain the knowledge of the NWS vs. IT feature. The benefits of LTN we want to evaluate is the ability to take into account information a neural network can not, giving the LTN an advantage. We acknowledge that the knowledge encoded in this example could (partly) be expressed as a feature in the dataset; however, real logic can express complex relationships that cannot be represented as features in the dataset. We have added a discussion about this in the paper, and state that we expect domain experts to specify more complex axioms.

Since this is a binary classification problem, the reported precision and F1 scores are poor. What are the state-of-the-art scores for this dataset in the literature? This comparison is essential.

- We have added a comparison of performance to two related papers on network intrusion detection using a similar dataset in the result section. We also emphasise the goal of the experiment: to show that a detector with domain knowledge performed better than a purely neural network under otherwise the same conditions. Here, we compare with two existing (and published) baseline detectors and show how our soultion with enriched with knowledge performs comparable or better for the two specific attack classes. We are comparing to these papers:
  - Jiyeon Kim, Yulim Shin, Eunjung Choi, et al. *An intrusion detection model based on a convolutional neural network.* Journal of Multimedia Information System, 6(4):165–172, 2019.
  - Arnaud Rosay, Florent Carlier, and Pascal Leroux. *Mlp4nids: An efficient mlp-based network intrusion detection for cicids2017 dataset. In Machine Learning for Networking*: Second IFIP TC 6 International Conference, MLN 2019, Paris, France, December 3–5, 2019, Revised Selected Papers 2, pages 240–254. Springer, 2020.

Section 4.2

It is recommended to use the same example in the prompt and in Figures 4, 5, and 6 for better clarity and coherence.

- To improve clarity the query-part of the prompt is now in accordance with the examples in the other figures.

Authors must discuss the tractability and complexity of Tlingo programs. Can this technique scale in real-world scenarios with hundreds of thousands of alarms?

- Address scaling by making the complexity of LTL/TEL and the intractability of Telingo explicit. We describe the limitations of experiment example and point to applicability of less expressive temporal approaches that have shown to be efficient and scalable in practice.

Section 4.3

In Figure 8, which labels are assigned to the classifier? It is not explained in the text.

- Clarified which labels are assigned to the classifier in the text.

Page 16, lines 29-39. These paragraphs must be rewritten, it is quite difficult to understand the ideas here exposed.

- Text has been rewritten to convey the ideas clearer.

Page 17, lines 1-2. What are the nodes and edges of these graphs (log and alarm graphs)? Some example is needed here.

- Simplified the explanation, and provided an example.

Page 19, again, authors should discuss scalability issues. In this small example, execution times are around 100s. Can this technique be applied to large-scale logs of alerts?

- We now discuss issues with complexity and scalability. We also mention possible mitigating techniques.

Section 4.4

This experiment is based on the previous work of authors in references [19] and [21]. Notice that reference [20] (unpublished) looks like the same as [21].

- The unpublished reference has been removed now as the paper has been accepted. Reference updated.

Page 20, lines 33-47, this part is unreadable. Sentences are not well written and there are new concepts that have not been explained before, so it is very difficult to understand this part.

- We have rewritten the section. Concepts are reintroduced with a brief summary of their meaning and purpose, and ross-referenced in the paper. Further reading is included in the references and footnotes.

Authors must explain what LM-1A, LM-1B, and DF-1A in order to understand the SWRL rules at page 23. In this page, there are also unreadable sentences because of their syntax.

- We have reworded and restructured this to improve readability. The identifiers now align with the rest of the material.

Section 4.5

This section is mainly a review of a previous paper [101] with little contribution in the context of this paper.

- We have clarified the new contributions in this paper (now presented as bullet points)
- We have also corrected spelling and grammatical errors and corrected inaccurate figure captions

Apart from these comments, the paper needs a careful reading for fixing numerous typos and syntax errors.

- We have proof-read the paper several times and improved the wording, corrected syntax errors and typos throughout the paper to enhance the quality.

Also the references need a lot of polish. There are many incomplete references, some duplicates and missing citations in the text.

- We have gone over all the references. Duplicate references have been removed, referenced arxiv papers properly, and performed general quality control.