We thank all the reviewers for the valuable feedback. We have modified the paper according to the reviewers to improve it.

Below we have mentioned the individual answers for each questions.

1 Review 1

Overall Impression: Good

Content: Technical Quality of the paper: Good Originality of the paper: Yes, but limited Adequacy of the bibliography: Yes, but see detailed comments

Presentation: Adequacy of the abstract: Yes Introduction: background and motivation: Good Organization of the paper: Satisfactory Level of English: Satisfactory Overall presentation: Good

Detailed Comments:

In this paper, the authors provide an overview of the existing literature on neuro-symbolic methods for trustworthiness, where this notion is seen as composed of five dimensions: interpretability, safety, robustness, fairness, and privacy.

A strength of the paper is its clarity, especially with regard to the methodological choices made for selecting the articles to consider in the survey.

Nonetheless, the work in its present form has some crucial weaknesses to be addressed:

- First and foremost, the article lacks of a proper discussion of the papers surveyed. This is usually done by providing a synthetic, though effective, description of each paper surveyed. Alternatively, the survey can highlight common perspectives, strategies, results, etc. among the papers. This is done in part in Sections 4.1 and 4.2. However, we believe that the level of detail of these sections is not sufficiently fine-grained to fully deliver the potential of a survey paper, where the reader should gain specific knowledge about the surveyed articles. → *added a full section to explain the main papers in one sentence (section 4.2: Description of the reviewed papers)*

- There seems to be an inconsistency in the information about the timeframe considered for the survey. Although it is initially declared that the work constitutes a systematic review of the recent literature from 2021 to 2022 (page 2 line 8), the authors then say that they "focused on

papers published in top academic venues from 2021 to 2023, including those available up to May 2023" (page 5 line 8). Moreover, Figure 1(a) is in line with the former statement and contains no information on 2023 papers. → *removed mention of 2023 but added a comment about the fact that it was the last advance at the time of writing*

– Finally, the general impression is that the actual scope of this survey is interpretability alone. As the authors themselves point out, "interpretability is the most extensively addressed aspect of trustworthiness" (page 3, line 24). The other four aspects – robustness, fairness, privacy, and safety – have a marginal discussion in the work. This is due to two different reasons. As far as robustness is concerned, the authors intentionally left out studies mainly centered on robustness, arguing that the concept is overly intertwined with that of accuracy. Regarding fairness, privacy, and safety instead, we have to wait until Section 4.4 to learn that there is a scarcity of neuro-symbolic applications for these topics (only one paper is mentioned, in relation to fairness). → *rephrased the title with interpretability and clarified this in introduction, abstract and survey results*

Questions:

– What justifies the methodological choice of focusing on proceedings papers only? Neuro-symbolic AI is a hybrid research topic, in the middle ground between computer science and logic, a field where journal papers have high relevance (e.g. the journal Neurosymbolic AI to mention one). → *Thank you for bringing that to our attention. To the best of our knowledge, the Journal of NeuroSymbolic AI had just launched around early 2023, which coincided with the time we were preparing our manuscript. At that stage, it appeared to be in its formative phase, and unfortunately, there was a limited number of published articles available for reference. While we are aware of other emerging venues such as those hosted by IOS Press and Sage, many of them were either very recent or not yet fully established at the time of our review. Additionally, although several well-regarded AI journals exist, their primary focus did not align specifically with the neuro-symbolic AI domain. As a result, we concentrated our literature review on top-tier AI conference proceedings, where substantial and timely research in this area was more readily accessible.*

– Why the following paper was excluded by the survey? Wagner, B. & d'Avila Garcez, A. S. (2021). Neural-symbolic integration for fairness in AI. In: CEUR Workshop Proceedings. AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021), 22-24 Mar 2021, California, USA. → *While this looks like a relevant paper for sure, unfortunately we didn't include any workshop papers in our survey, to keep the number of papers included reasonable and ensure we are focusing on top-tier papers.*

Finally, we suggest some minors improvements to the text:

– Legend of Figure 1(b) contains the typo "where taken"; → *changed*

– Figure 3: we suggest changing the abbreviation "AMB" for "ambiguous" with a more self-explanatory expression - "ambiguous" or "unclear" should be fine. → *changed*

=== final meta-level thoughts ==

I think it would be good to have some sort of comparison to existing works that discuss XAI and trustworthiness from a non NeSy AI context, and particularly to human modelling, and discussions on well-established surveys and experimental evaluations (e.g., stakeholder discussion cf Adrian Weller, Trustworthiness and XAI taxonomy by Nathalie Rodriguez and colleagues). → *We have added references and discussions of these works in the related works section. For the comparative analysis, we ensured that the cited studies support the key dimensions of our proposed taxonomy. We tried our best to align the work with existing type of classifications from research, but from the executive point of view it may lack some degree.*

There is an attempt to classify existing works which is appreciated but further work is needed in terms of updating the reader on non NesyAI solutions. Therefore, how well are NeSy AI solutions addressing these concerns? What's hard to do (see work on XAI planning and explaining in dynamic domains). → *added a subsection 6.3: Common Challenges*

2 Review 2

Overall Impression: Average

Content: Technical Quality of the paper: Average Originality of the paper: Yes, but limited Adequacy of the bibliography: Yes

Presentation: Adequacy of the abstract: Yes Introduction: background and motivation: Good Organization of the paper: Needs improvement Level of English: Satisfactory Overall presentation: Good

Detailed Comments:

This paper focuses on applying Neuro-Symbolic (NeSy) methods to interpret deep learning systems. It proposes a novel, comprehensive, and multi-perspective framework that systematically categorizes and analyzes recent literature in this field.

Reasons to accept: 1. The paper categorizes existing articles reasonably, integrating NeSy methods to increase the trustworthiness of deep learning systems. 2. The paper defines different types of NeSy methods mainly from two perspectives, Symbolic Data Structures and types of interpretability. 3. The proposed categories proposed are clear and move toward addressing the problem of unclear categorization in NeSy field. 4. Although the paper lacks quantitative analysis, it does keep the discussion from going deeper and more reliable. It doesn't really influence the framework they proposed.

Reasons to reject: 1. There is insufficient forward-looking discussion. Although the paper mentions future research directions, it may lack sufficient innovative discussion or foresight. In particular, the paper points out the lack of NeSy applications in privacy and fairness, but provides no further insight into it. → *added more development*

2. The paper lacks a quantitative analysis, which keeps the discussion superficial. There also seems to be a lack of listing current metrics or standards to measure how the NeSy model enhances model interpretability. → *added discussion of the fact that interpretability is too often self-assessed and there is a lack of current metrics*

3. The writing style is wordy and there are many internal repetitions. For example, in Section 3, the paper mentions they "selected the papers directly contributing to trustworthiness" three times. → *we improved the phrasing of some sentences*

4. The paper suffers from unclear citations. In Section 1.2, line 43, the paper mentioned "some systems", but no clear reference can be found to prove the argument. → *rephrased to make it clear* Specific references should be provided.

5. Time conflict. In Section 1, the paper notes that the studied article is from 2021-2022, but in Section 3 they are mentioned from 2021-2023. The conflict may confuse the audience. → *Thanks, fixed as mentioned in the comment of reviewer 1.*

6. In Section 3, the paper claims they have surveyed the papers from AACM FAccT, but Fig.1(b) shows there is 0 paper coming from AACM FAccT. The paper may point out less content related to NeSy for Trustworthy in AACM FAccT, but fails to demonstrate it explicitly. → *Some comments about that are now in section 3.3: Some Statistics*

7. Consider adding an overview paragraph to make it easier for purposeful readers to quickly locate the section they need. → *done*

8. The paper could describe the content of Table 1 more clearly in the caption. → *added reference to the section explaining the categories*

3 Decision Letter

Decision Letter: Thank you for your submission to Neurosymbolic Artificial Intelligence.

This is to inform you that based on the reviewer's comments, your paper requires major revisions. Please carefully take into account the enclosed comments by the reviewers when preparing the revised version. It is incumbent upon you to do so. Please provide punctual responses to the issues raised by the reviewers and prepare a separate text file containing such responses.

Please pay special attention to the following in your revision:

Provide a description of each paper surveyed or highlight common perspectives, strategies, results, etc. among the papers in a greater level of detail than currently written in Sections 4.1 and 4.2. [**?**] Address the focus on interpretability by either revising the scope / focus of the work (with appropriate edits to the title, introduction, etc.) or expand discussion of robustness, fairness, privacy, and safety. If any are omitted from consideration, clarify this as early as possible in the text. → *The central role given to interpretability has been clarified and stated in the title, and the only discussed concept which has been excluded, robustness, is now mentionned only in the background where we explain why we excluded it.*

Ensure that sources are cited clearly, i.e., avoid oblique references (such as "some systems") pointed out by reviewer #2. → *made the references more precise*

Methodology Clarify the date range of reviewed literature → *done* Clarify/expand your justification for excluding all journal venues → *Mentioned the justification in the review feedback of reviewer 1 and also added in the justification in the paper (section 3.1 Methodology)*

Explain why AACM FAccT was considered in scope for the review but evidently yielded no content which was included in the review → *added a short explanation (in section 3.3: Some Statistics)*