Response to Reviewers

Manuscript Title:
Neurosymbolic Architectures for Algorithmic Fairness

Manuscript Number: *#894-1905*

Dear Editors of the Special Issue on *Trustworthy Neurosymbolic AI in Regulated Domains*,

We appreciate the opportunity to revise our manuscript based on the reviewers' comments and thank them for their feedback and suggestions. We summarize the main revisions made to the manuscript below, followed by our point-by-point response to each reviewer.

1. We carefully revised the introduction section to better substantiate our arguments and updated the contribution subsection to precisely articulate the type and objective of our paper.
2. We revised and re-structured both the "Bias Mitigation" (now entitled "'Bias Mitigation from a Neurosymbolic Perspective", with new paragraphs on existing neurosymbolic techniques in bias mitigation) and "Neurosymbolic Architectures" (now integrated as a preface into "Neurosymbolic Architectures for Bias Mitigation") sections to better integrate these overviews into the main storyline of the paper and clearly separate them from existing surveys.
3. We added more recent approaches and references to the "Neurosymbolic Architectures for Bias Mitigation" section (e.g., ProbLog4Fairness; Adriaensen et al. 2026).
4. We added an "Illustration" section, in which we ran an experiment to showcase and discuss the differences between one particular neurosymbolic approach (LTN) and a similar non-neurosymbolic approach.
5. We added an Appendix with detailed information about our literature research methodology.

We believe the revised paper significantly improved in clarity and its contribution and will provide a useful resource and reference for the emerging research at the intersection of Neurosymbolic AI and algorithmic fairness.

Reviewer #1

- Reviewer's comment: *This research provides a systematic mapping between neurosymbolic architectures and stages of bias mitigation. The topic is timely and a valuable contribution, as the integration of symbolic reasoning into fairness-aware AI has not been sufficiently explored. With improved structure, clearer differentiation from existing surveys, and enhanced presentation of figures and references, the work is publishable. The introduction appears to lack sufficient references and could benefit from clearer phrasing in several parts. The Bias Mitigation and Neurosymbolic Architectures sections show considerable overlap with previous studies, and the differences should be made more explicit. The remaining sections are suggested for minor revisions.*

  Response: We thank Reviewer #1 for their comments and helpful feedback. We made significant changes to all mentioned sections to satisfy these remarks. In the introduction, we added references and clarified our claims to provide a clear and well-supported foundation for our contribution. In the "Bias Mitigation" section ("Bias Mitigation from a Neurosymbolic Perspective" in the revised manuscript), we added several paragraphs to discuss existing techniques from a neurosymbolic perspective. Furthermore, we shortened the "Neurosymbolic Architectures" section and integrated it as a preface into the "Neurosymbolic Architectures for Bias Mitigation" section. With these changes, we hope to improve the storyline of the article and mark a clearer separation from other studies. We elaborate on these revisions below.

- Reviewer's comment: *1. Missing references:*

a *Trustworthiness aspects*
Are these aspects your own interpretation? Or should they be cited?

b *"Researchers in the field of neurosymbolic AI have proposed numerous architectures that incorporate the understandable, reasonable nature of symbols and statistical models that can handle noise and uncertainty."*
It is expressed as if there are clearly past studies, such as "proposed numerous architectures", yet no such references are provided.

c *"Most bias mitigation approaches encode constraints directly into the machine learning procedure and thus implement specific fairness notions for a distinct set of use cases."*
Add the case as a reference.

d *"The biggest issue in symbolic systems is the grounding problem, i.e. to find an adequate mapping between the continuous real world and the assumed discrete world of the model".*
This is a well-known issue that has been the subject of past research, so a citation should be included. Harnad, Stevan. "The symbol grounding problem." Physica D: Nonlinear Phenomena 42.1-3 (1990): 335-346.

Response: Thank you for raising these points. We have added references to the sentences referred to in (a), (b) and (d) to substantiate our arguments. We have re-phrased sentence (c) as the essential point of the argument here is simply to highlight the diversity of existing fairness notions.

- Reviewer's comment: *2. Ambiguity and suggested rephrasing*

a *"Fairness stands out among these concepts, as this is an aspect of trustworthiness that ADM actually promises to enhance compared to human decisions".*
Clarify how fairness differs from the other components of trustworthiness, and in what way ADM systems are assumed to improve fairness relative to human judgment.

b *"While bias detection queries whether data or a prediction satisfies a fairness constraint, bias mitigation employs fairness constraints on the data, the prediction model or the output".*
Does this mean "While bias detection queries whether predictions satisfy a fairness constraint, bias mitigation employs fairness constraints on the data and the models"?

c *"i.e., are tied to one single formal definition of fairness"*
What is meant by "one single formal definition"?

Response: We have revised all of these passages to increase the precision of our argument. This includes (a) distinguishing fairness from the other components of trustworthiness and clarifying its role in ADM systems, (b) clarifying the difference between bias detection and bias mitigation as suggested and (c) clarifying the link between bias mitigation techniques and fairness notions (with examples).

- Reviewer's comment: *3. Clearly state the distinctions from other review studies in the same field. emphasise that this paper maps types of neurosymbolic approaches to bias mitigation techniques. This distinction should be explicitly linked to your listed contributions.*

Response: We restructured and extended the overview sections to mark a clearer separation between our work and prior studies. In our contribution paragraph, we added stronger emphasis on the purpose and novelty of our work and the distinctions from other review studies.

- Reviewer's comment: *4. Clearly state the ultimate outcome or impact of the listed items in the contribution subsection.*

Response: We have expanded the contribution subsection to clearly state the intended impact of our work, which we see as a critical starting point to connect the currently disjoint domains of algorithmic fairness and Neurosymbolic AI with the ultimate goal of developing novel neurosymbolic bias mitigation methods.

- Reviewer's comment: *1. Missing reference on "Many wide-spread notions of fairness focus on binary classification tasks with one binary protected attribute". For example: Pagano, Tiago P., et al. "Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods." Big data and cognitive computing 7.1 (2023): 15.*
2. Explain "b" in the equation in Group Fairness

> **Response:** Thank you. (*1*) We have added the suggested reference and (*2*) revised our notation in section 2.2 to consistently use a and a' to denote different groups defined by protected attributes.

- Reviewer's comment: *1. Although well researched, the presentation is overly enumerative and challenging to follow, with unclear transitions to the subsequent section. The process of bias mitigation has already been reviewed in: Hort, Max, et al. "Bias mitigation for machine learning classifiers: A comprehensive survey." ACM Journal on Responsible Computing 1.2 (2024): 1-52. Rather than merely introducing it, please clearly articulate how your perspective differs from or extends it.*

> **Response:** We agree that it is important to clearly link this section to the main objective of the paper. We have considerably revised our overview of bias mitigation techniques (now titled "Bias Mitigation from a Neurosymbolic Perspective") and added new subsections that discuss intersections with the field of Neurosymbolic AI. Additionally, we shortened some parts of this section that are not relevant for this paper later on. Thus, our (revised) overview is explicitly tailored to provide connection points for readers in Neurosymbolic AI, which we now clearly articulate at the beginning of this section (page 8).

- Reviewer's comment: *1. While new references have been added, the content appears largely identical to the following research: Kautz, Henry. "The third ai summer: Aaai robert s. engelmore memorial lecture." Ai magazine 43.1 (2022): 105-125 Emphasize the distinct perspective of this manuscript.*

> **Response:** We have integrated this section into the following section ("Neurosymbolic Architectures for Bias Mitigation") and shortened it. We purposefully chose the taxonomy of Kautz (2022) as it provides a well-known structure for readers in Neurosymbolic AI that we can use for our mapping between neurosymbolic architectures and bias mitigation methods.

- Reviewer's comment: *1. "Chiappa (2019) proposed a method to adjust the output of a predictor to satisfy counterfactual fairness." However, this is not listed in the references from Table 1.*

> **Response:** Thank you for the note. We re-evaluated this paper and moved it to the "Neuro[Symbolic] Bias Mitigation"-subsection, next to including it in Table 1.

- Reviewer's comment: *2. Adding labels to each object in Figure 2 would be preferable to. The first model from left is SCM, the second is the neural network, and the third is the generated data. This is not explained in Section 5.3 (Neuro:Symbolic→Neuro Bias Mitigation).*
3. Figure 3 would also be preferable to add labels to each object.

> **Response:** Thanks, we added labels to each object in both figures (Figure 1 and 3 in the revised manuscript) to improve clarity. We have also extended the caption of Figure 1 and added another explaining sentence to the respective section.

Reviewer #2

- Reviewer's comment: *This paper presents a well-structured and timely survey that maps neuro symbolic architectures to different stages of bias mitigation and the overall framing is sensible, with the taxonomy being useful for understanding how symbolic reasoning might support fairness interventions at various points in the ML pipeline. The paper fits well within the scope of the special issue and has potential*

as a reference or entry point for researchers working in this domain. That being said, several aspects would benefit from clarification / tightening before publication.

> **Response:** We thank Reviewer #2 for their positive assessment. We have carefully addressed all raised issues as detailed below.

- Reviewer's comment: *A number of strong or normative claims in the introduction would benefit from clearer grounding or qualification. These include such statements as suggesting automated decision-making systems "promise" improved fairness compared to human decision-making, or that "most" bias mitigation approaches encode constraints directly into learning procedures. These claims are central to the paper's motivation but currently under-supported. Narrowing the scope or adding references would strengthen the framing and avoid over-generalisation.*

> **Response:** We have carefully revised the introduction to improve the precision and support of our arguments (see also our response to Reviewer #1 above). This includes elaborating on the comparison between ADM and human judgment and adding references concerning human moral judgment. Also, we relaxed the claim that "most bias mitigation methods encode constraints directly into the learning procedure" and now merely rather note the diversity of existing fairness definitions (see page 2).

- Reviewer's comment: *Relatedly, while the paper's main contribution is conceptual rather than technical, the novelty of the proposed mapping could be made more explicit. Several sections overlap in spirit with existing surveys on bias mitigation and neurosymbolic AI, and it would help to more clearly articulate how this work differs from or extends those prior efforts, particularly in relation to previously cited neurosymbolic fairness approaches.*

> **Response:** We have considerably revised and restructured our overview sections (see also our response to Reviewer #1 above), next to more clearly articulating how our paper differs from related prior work in the contribution subsection (pages 4–5).

- Reviewer's comment: *The treatment of fairness notions is generally accurate but remains quite high-level. Given the focus of the special issue, clearer signposting of which fairness definitions are assumed or in scope (e.g. group vs individual, statistical vs causal) and brief acknowledgement of known trade-offs would strengthen the conceptual rigour without significantly expanding the paper.*

> **Response:** We have clarified the role and scope of section 2 (now labeled "Recap: Algorithmic Fairness") in the revised manuscript. This includes clear signposting of which categories and aspects of algorithmic fairness are (not) considered early on and a concise discussion of known impossibilities (page 5).

- Reviewer's comment: *From a presentation perspective, the taxonomy sections are informative but clearer transitions or short comparative summaries would improve narrative flow (vs the current enumerated list).*

> **Response:** We have integrated the taxonomy of neurosymbolic architectures into the following section ("Neurosymbolic Architectures for Bias Mitigation") and clarified its role to improve the narrative flow of the paper and clarify that its main purpose is merely to provide a structure for the following mapping between neurosymbolic architectures and bias mitigation approaches (see also response to Reviewer #1 above).

- Reviewer's comment: *Finally, there are a small number of minor typographical and notation inconsistencies (e.g. variable naming in some fairness definitions on page 5). These appear to be local and non-conceptual, but should be corrected for clarity and consistency.*

Response: Thank you for raising this. We have streamlined the presentation of fairness definitions and resolved notation inconsistencies (i.e,, we now use a and a' to denote groups defined by protected attributes throughout). Note that "groups" in multi-group fairness may not be discrete and thus we use a different notation in this subsection.

- Reviewer's comment: *Overall, this is a solid and relevant survey contribution and with clearer grounding of the key claims, a more explicit articulation of the conceptual contribution, and some small improvements to the presentation and precision, the paper would be suitable for publication.*

Reviewer #3

- Reviewer's comment: *The following comments can be addressed: 1-The paper is almost entirely conceptual and does not include any empirical evaluation to demonstrate that the proposed architectures work in practice.*

Response: We thank Reviewer #3 for their comments and assessment of our work. The main approach of our paper is indeed conceptual – it is not meant to be an empirical study or a classic literature survey. We rather see our work as a *Perspective* paper, similar to, e.g., the article by Wagner and d'Avlia Garcez (2025) (published in *Neurosymbolic Artificial Intelligence*). Yet, we added an illustration section in which we compare one of the discussed neurosymbolic approaches to a baseline and a non-neurosymbolic counterpart (section 5). Additionally, we added a methodology section to the Appendix to legitimize our claim of a complete summary of existing neurosymbolic bias mitigation methods.

- Reviewer's comment: *2- No experimental comparison is provided against existing fairness-aware neural or neurosymbolic methods, making it impossible to assess practical benefit.*

Response: We agree that this kind of comparison was missing in the paper. In our newly added illustration section, we compare an LTN with a counterfactual fairness axiom to a very similar non-neurosymbolic model. On this basis, we discuss that the promises of this particular approach are rather qualitative than quantitative.

- Reviewer's comment: *3- The authors should include at least one implemented architecture with quantitative results on standard fairness benchmarks (for example, COMPAS or Adult). Similar to the following paper, you could provide neumerical results and compare it with 3 existing approaches you mentioned like the following one. They did it on Adult, COMPAS and Lawschool datasets. Heilmann, X., Manganini, C., Cerrato, M., & Belle, V. (2025). A neurosymbolic approach to counterfactual fairness. In 19th International Conference on Neurosymbolic Learning and Reasoning.* `https://openreview.net/pdf?id=YZSDHz3Ydb`

Response: As mentioned above, we now provide this kind of comparison in the illustration section. We built on the work of Heilmann et al. (2025) but use the ACSPublicCoverage dataset for our experiment. Our experiment further differs from Heilmann et al. (2025) in that our compared non-neurosymbolic architecture only differs from the LTN in the learning objective. Thereby, we aim to show the particular difference between purely statistical and neurosymbolic approaches to implementing constraints in loss functions.

- Reviewer's comment: *4- If the paper is intended as a survey, it lacks the depth, systematic methodology, and coverage expected from a review article.*

Response: Thanks for raising this point. While our paper is not primarily intended as a survey, we added a survey methodology section in the Appendix. This section supports our claim that our summary is comprehensive in the sense that it fully covers the small number of existing studies on neurosymbolic bias mitigation.

- Reviewer's comment: *5- The proposed taxonomy in Table 1 is nice, but it should be validated empirically.*

> **Response:** We understand the demand for an empirical validation of this table. However, we do not see this within the scope of this paper – Table 1 is supposed to present and qualitatively compare the neurosymbolic approaches discussed in section 4 and give them some sort of classification. This conceptual mapping aligns with our main goal of providing researchers from neurosymbolic AI and algorithmic fairness a perspective for future research at the intersection of both domains. Thus, we do not make normative claims about single approaches, but rather want to highlight promises of this broader class of methods.

- Reviewer's comment: *6- The paper does not discuss how sensitive attributes are obtained or inferred in practice, which is critical for fairness systems.*

> **Response:** Thank you for raising this important aspect. We have clarified our operating assumption regarding the availability of protected attributes and integrated research on fairness assessments under unawareness of such attributes in the revised section 2 (page 5).

- Reviewer's comment: *7- The paper would benefit from a comparison table summarizing existing fairness methods versus proposed neurosymbolic approaches in terms of guarantees, cost, interpretability, and accuracy.*

> **Response:** Thank you for this idea, we generally agree that such comparisons would help the reader. In light of our main objective and our conceptual contribution (see above), we added a *Promises* column to Table 1.

- Reviewer's comment: *8- To be publishable as a regular research paper, the authors should add empirical validation, structured benchmarking, and reproducible experimental protocols; otherwise, the work should be reframed and significantly expanded as a systematic survey.*

> **Response:** As outlined above, our paper primarily offers a conceptual perspective rather than providing a literature review or an empirical work. Yet, we acknowledge the concern and extended our work with an illustration section, next to clarifying our survey methodology and considerably extending and revising our literature overview and mapping sections.

References

Adriaensen R, Van Praet L, Bekker J, Manhaeve R, Delobelle P and Buyl M (2026) Problog4fairness: A neurosymbolic approach to modeling and mitigating bias. *Proceedings of the AAAI Conference on Artificial Intelligence* 40(24): 19542–19550. DOI:10.1609/aaai.v40i24.39033. URL `https://ojs.aaai.org/index.php/AAAI/article/view/39033`.

Chiappa S (2019) Path-specific counterfactual fairness. In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, pp. 7801–7808. DOI:10.1609/AAAI.V33I01.33017801. URL `https://doi.org/10.1609/aaai.v33i01.33017801`.

Heilmann X, Manganini C, Cerrato M and Belle V (2025) A neurosymbolic approach to counterfactual fairness. In: *19th International Conference on Neurosymbolic Learning and Reasoning*. URL `https://openreview.net/forum?id=YZSDHz3Ydb`.

Wagner BJ and d'Avlia Garcez A (2025) A neurosymbolic approach to ai alignment. *Neurosymbolic Artificial Intelligence* 0(0): NAI–240729. DOI:10.3233/NAI-240729. URL `https://doi.org/10.3233/NAI-240729`.