Response to Reviewers

Manuscript Title:
Neurosymbolic Architectures for Algorithmic Fairness

Manuscript Number: *#962-1991*

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

Thank you for this second revision opportunity and thanks to the reviewers for their further feedback. Please find our point-by-point response to the remaining reviewer's remarks below. With the implemented adjustments, we hope to have sharpened the profile and readability of our article.

Reviewer #1

- Overall. This manuscript is a comprehensive and narrative survey on neurosymbolic AI for fairness, and it is particularly interesting for its discussion of the relationship between bias mitigation and neurosymbolic methods. This work addresses an important gap: although several studies have applied neurosymbolic approaches to fairness problems, it remains unclear how logical rules and fairness constraints can be integrated into neural networks and what kinds of effects these approaches provide. Section 4 clearly explains the relationship between neurosymbolic methods and bias mitigation, and Section 5 demonstrates the characteristics of the neurosymbolic approach through experiments. On the other hand, the introduction and Sections 3 would benefit from clearer positioning of the paper's main motivation and consistency.

  > **Response:** Thank you for your positive assessment and the encouraging words. We have revised the introduction and Section 3 as outlined below to further improve the positioning and consistency of the paper.

- Introduction. While the introduction presents ADM primarily in the context of public policy, it remains somewhat unclear whether the paper focuses specifically on neurosymbolic fairness models for public services. In addition, the frequent use of supplementary explanations and examples in parentheses interrupts the argumentative flow and reduces readability.

  > **Response:** Thank you for these comments. We added a clarifying sentence to the introduction paragraph *Fairness in Automated Decision Making*, in which we clarify the scope of our overview and that public policy is primarily used as a motivational (but not limiting) framing for the methods introduced and discussed in the paper.
  > We have further removed or revised additional explanations that were originally given in parentheses to improve readability.

- Section 3 (Bias Mitigation from a Neurosymbolic Perspective). A substantial portion of the paper is devoted to reviewing conventional bias mitigation techniques in pre-processing, in-processing, and post-processing. As a result, the central novelty of the paper, the neurosymbolic perspective on fairness and bias mitigation, becomes less prominent. Condensing the general survey and allocating more space to neurosymbolic-specific architectures, constraint representations, and reasoning mechanisms would strengthen the paper's contribution and better highlight its originality. This would also help distinguish the paper more clearly from existing survey papers on general fairness and bias mitigation techniques for classification models.

Response: We have further condensed the *Common Approaches* in Section 3 (*Bias Mitigation from a Neurosymbolic Perspective*) by removing references and explanations that are not necessary for the neurosymbolic perspective. The current content of the *Common Approaches* subsections now concisely provides the background for the following elaborations on *Neurosymbolic Pre-, In-, and Post-Processing*, which we have expanded as suggested to strengthen the papers specific positioning and contribution. These sections still provide references as entry points to the more extensive collection and description of neurosymbolic architectures for bias mitigation in Section 4.