# NAI Review Revision

Benedikt Wagner

Jan 2024

We thank the reviewers for their valuable feedback and contributions. We have rewritten the entire introduction, added more structure to the background assumptions, highlighted claims, and focused the introduction of several terms (which were previously buried inside the introduction). In the following we will outline the major revisions made to the original script that specifically address the reviewers comments.

## 1 Reviewer 1

Reviewer comments followed by our changes:

- To clearly articulate in a few highlighted lines what the key claim or position is that the authors want to make or defend:

  Incorporated several precise statements in lines 34-38 to directly address the key claims and our position while being aligned with the contributions identified by reviewer 1. The abstract and introduction of the paper (lines 13-39) clearly articulate the key claims and positions of the authors. The emphasis is on neurosymbolic integration for AI alignment with a focus on concept-based model explanations, offering AI systems the ability to learn from human revision and assisting humans in evaluating AI capabilities.

- to make and highlight the assumptions explicitly – like that the predicates / concepts are what makes:

  The assumptions underlying the neurosymbolic concept integration are explicitly stated in section 2 (lines 47-50), outlining the fundamental role of predicates and concepts in the structure and functionality of neurosymbolic AI systems.

- to remove those connections that are not essential to this position, e.g. the communication channels, do you really need FOL here ?:

  Significantly shortened the references to communication theory and clarified the reference to FOL in (new) section 3

- to clearly define or introduce concepts such as grounding (which seems to be different than symbol grounding in an interactive environment), concepts, explanations, semantic representation space:

  The paper clarifies terms like grounding, concepts, explanations, and semantic representation space in section 4 (lines 26-41). This section provides definitions and contextualises these terms within the neurosymbolic AI model.

- to very briefly illustrate what LTNs do (to make the position paper stand by itself); what are the inputs/outputs?

  LTNs and their functionality are described in section 4.1 (lines 45-39). This includes an explanation of how LTNs process and reason about complex symbolic logic statements using neural network architectures and the integration of symbolic knowledge in the form of logical predicates or rules. Extensively elaborated on inputs/outputs/groundings etc.

- (what is a Broden set, what are the members of G(logic) ... that is not said, are these images, examples

  Expanded the description of the Broden set and its contents, along with a more detailed discussion on grounding.

- to discuss whether alternative systems than LTN could be used.

  Section 4 now acknowledges that LTNs represent just one of several methods capable of merging symbolic logic with neural networks, mentioning alternatives that employ differential logic-based loss functions for neural networks to adhere to logical constraints while learning from data (p.6 lines 33-39).

- to introduce a more detailed and more precise example of what is done and why, at the start of the paper

  This is hopefully clear now with all of the changes in the introduction/changes splitting of first sections. We

- The fairness needs more explanation (like what is RMI and RFI).

  Detailed explanations regarding fairness, including RMI and RFI, are now included in the relevant (fairness) section.

- Also, what do fairness and explanation have in common ? And are these the only things that matter for value alignment. It still seems a big step to go from these two to value alignment. This is insufficiently argued.

  This relationship is now thoroughly defined in the discussion, with an expanded introduction in the background section. The paper discusses how integrating knowledge extraction from deep networks into the LTN framework and adding tailored fairness constraints can instill fairness into deep networks (lines 44-51). The same mechanism can be used to explain and revise any model using concept representations.

- to connect to recent work on concept-based and explainalbe neurosymbolic systems, there has been a whole line of research on this that is not mentioned here (on concept embedding models and related)

  A large number of recent work is now addressed in the latter part of section 2 and beginning of section 3.

## 2 Reviewer 2

- In response to your point on positioning, we have clarified our stance on the AI alignment problem. Our revised manuscript underscores the utilization of neurosymbolic AI systems for aligning AI with human values, emphasizing the pivotal role of symbolic representations in enhancing the explainability and transparency of decision-making processes in AI (Section 1). This approach sets our work apart from purely reasoning focused NeSy methods and underlines the criticality of interactive explainability in AI systems, which we believe is foundational for value alignment (Sections 1 and 2).

- Vagueness: In the original manuscript, the introduction and discussion were broad and lacked direct focus on the AI alignment problem. We have now revised these sections to provide a clearer and more direct exposition of our thesis and argument.

  - Introduction and Focus on AI Alignment Problem: Original: Discussed neurosymbolic AI in a general sense without clear connection to AI alignment (original, Sec. 1). Revised: Directly connect neurosymbolic AI's role in AI alignment, emphasizing its importance in ensuring fairness and providing explanations as foundational elements for value alignment (Introduction)
  - Technical Detail and Relevance to AI Alignment: Original: Sections on technical details (e.g., logical language use, model querying) did not clearly tie back to the alignment problem (Sec. 4). Revised: Refined the discussion of technical aspects like model querying and logical language use, explicitly linking them to how they facilitate AI alignment and fairness ( Sec. 4).

- Novelty: We have now more clearly delineated our position and how it differs from and contributes to existing approaches in AI alignment:

  - Distinct Approach to AI Alignment: Original: Presented our approach without clearly differentiating it from others (Introduction). Revised: Clarified how our approach, focusing on fairness and explainability through neurosymbolic AI, offers a unique contribution to the field.

  - Scientific Interest and Novelty: Original: Lacked explicit discussion of the novelty and scientific interest of our approach (Sec. 1) Revised: Explicitly highlighted the novelty and scientific relevance of integrating fairness constraints and explainability into AI alignment using our neurosymbolic approach (Sec. 2 and 4).
  - Existing literature: Original: Lacked the direct comparison to relevant literature. Revised: Added references and direct comparisons to all of the mentioned papers (Sec. 2)

- Writing: The writing style and structure have been improved for better clarity and flow:
  - Structural Clarity: Original: The introduction appeared more as a chain of thoughts with less structured flow (Introduction). Revised: Restructured the introduction and subsequent sections for a more conventional and coherent scientific exposition (Introduction and Sec. 4).
  - Connection Between Sections: Original: Abrupt shifts in sections from historical arguments to technical details (nai-paper-688.pdf, Sec. 1 and 2). Revised: Ensured smoother transitions between sections, maintaining a consistent focus on AI alignment and the role of neurosymbolic AI (Sec. 1, 2, and 4).

# 3 Reviewer 3

- Citing Kambhampati et al. (AAAI 2022)

  We have acknowledged the significant contributions of Kambhampati et al. (AAAI 2022) in our discussion, highlighting the alignment of their insights with our neurosymbolic approach, particularly in the use of symbols for effective human-AI communication

- Concept-Based Models

  We have incorporated references to concept-based models in the background section, aligning with recent literature to underscore their relevance in creating an interpretable neurosymbolic pipeline.

- Interactive Adaptation Through Explanations

  We have highlighted our framework's compatibility with the principles of interactive adaptation, in line with the overview by Teso et al. (Frontiers in AI 2023), emphasizing user-guided explanations for model refinement.

- Addressing Concept Leakage and Reasoning Shortcuts

  We have added a new paragraph addressing the concern of reasoning shortcuts in neurosymbolic models, as highlighted by Marconato et al. This addition directly addresses the potential for semantic misalignment

in NeSy models. We elaborate on how our approach may mitigates the risk of reasoning shortcuts by ensuring semantic alignment between the AI's learned concepts and human logic, enhancing the model's interpretability and alignment with user expectations.