Review Response

March 17, 2026

Dear NAI Editor and Reviewers:

We thank the reviewers for their valuable time, careful reading, and insightful suggestions. The feedback helped us further improve the clarity and positioning of the paper. Below, we address each reviewer's comments directly.

Reviewer 1

We appreciate the reviewer's recognition that Sections 4 and 5 on NeuPSL and learning contain clear, significant technical contributions and that the experimental evaluation in Section 6 is thorough.

- **Notation Heaviness** We agree that notation could be reduced to focus on the primary contributions of the paper. In response, we made significant changes to Section 3, *A Mathematical Framework for NeSy*, by removing the definitions and notation for alternative inference tasks so that it stays focused on the central prediction setting. Additionally, we revised the *Modeling Paradigms for NeSy* subsection by moving the formal definitions to the appendix and removing the more formalized portions of the main-text examples, making Section 3 significantly more concise and reducing notation.

- **Value of the Mathematical Framework** We retained the explicit $\mathbf{g}_{nn}/g_{sy}$ distinction because it is doing more than introducing notation. Practically, it reflects the compositional design of real neural-symbolic systems. Theoretically, the proof of Theorem 6 leverages this framework to derive gradients of optimal value functions for neural-symbolic systems with general compositional structure. Without defining g_{sy}, that distinction and neural symbolic connection would need to be re-derived for specific systems rather than stated as a consequence of the framework's compositional structure. In this sense, the abstraction has both practical and theoretical value: it reflects modular NeSy design and enables general gradient results. We now make this explicit in the *Neural Symbolic Energy-Based Models* subsection of 3 immediately after Definition 1.

- **DSVar vs DSPar.** We now state this directly in *Modeling Paradigms for NeSy* of Section 3, "in DSVar the neural component predicts a subset of the target variables and these predictions are treated in practice as fixed evidence, whereas in DSPar the neural predictions parameterize the symbolic component and inference still optimizes over all target variables". That subsection also now connects this distinction to prediction, learning, and the resulting continuity properties and gradient forms.

- **Semantic Loss vs LTN categorization.** We clarified the scope of this discussion in *Modeling Paradigms for NeSy* of Section 3 and in the appendix subsection *Expressing NeSy Approaches via NeSy-EBMs*, so that the mappings are read as NeSy-EBM views of particular instantiations rather than as claims about the full expressivity of each framework. The intended distinction is based on the role the neural outputs play in the prediction program, not merely on the fact that both approaches ultimately backpropagate through differentiable objectives.

- **New contributions relative to the original NeuPSL paper.** We agree that this should be further emphasized and clarified. In the revised *Introduction* (Section 1), *A Mathematical Framework for NeSy* (Section 3), and *A Suite of Learning Techniques for NeSy* (Section 5), we frame the paper not only based on the NeuPSL instantiation, but also the broader NeSy-EBM framework, the modeling paradigms, and the learning suite built on top of that framework.

Reviewer 2

We appreciate the reviewer's assessment that the strongest contributions are the NeuPSL inference formulation and the learning suite, particularly the bilevel/value-function approach, as well as the recognition that the empirical evidence supports the core value proposition.

Major Revision Items

- **Streamline the NeSy EBM Framework Exposition (Section 3).** We agree with the reviewer's point that Section 3 introduced multiple abstraction layers that may not be necessary for the paper. In response, as mentioned in response to Reviewer 1, we streamlined the *Neural Symbolic Energy-Based Models* and *Modeling Paradigms for NeSy* subsections by moving the formal modeling-paradigm definitions to the appendix, removing the more formalized parts of the examples in the main text, and removing alternative inference tasks so that the discussion stays focused on the central prediction setting and the main components of the NeSy-EBM framework.
- **Clarify DSVar vs DSPar by inference topology, not argument placement.** We agree that this is the primary difference. We now state this directly in *Modeling Paradigms for NeSy*, "in DSVar the neural predictions are treated as fixed evidence, so the neural component fixes part of the inference decision variables, whereas in DSPar the neural predictions parameterize the symbolic component while inference still optimizes over all target variables." We also added a short practical guideline that DSVar is natural when a simpler prediction program or structured loss is preferred, while DSPar is natural when symbolic inference should be able to correct weak neural predictions.
- **Strengthen empirical positioning relative to external NeSy baselines (or narrow claims).** We clarified the empirical claims in the revised *Introduction* (Section 1) and *Empirical Analysis* (Section 6), and now make the scope explicit again in *Limitations* (Section 7), so that the results are presented as evidence for NeSy-EBMs as instantiated through NeuPSL rather than as an exhaustive cross-framework benchmark. For a broader cross-framework comparison context, we point readers to the comparison-oriented NeSy surveys and taxonomies already discussed in related work, such as Besold et al. (2022), De Raedt et al. (2020), Giunchiglia et al. (2022), and van Krieken et al. (2022), alongside prior NeuPSL papers reporting additional applications and analyses Dickens (2024); Dickens et al. (2024a,b); Pryor (2024); Pryor et al. (2023).
- **Expand discussion of ϵ-regularization.** We added a short discussion of ϵ in *A Smooth Formulation of Deep HL-MRF Inference* and point readers there, as well as to the appendix, for the empirical analysis. The revised text now states that $\epsilon \geq 0$ is added to ensure strong convexity, that the effect on prediction performance observed in practice is small in the range from 0.01 to 10, and that increasing ϵ can substantially decrease inference runtime. We also introduced a D-BCD inference algorithm from Dickens et al. (2024a) and the sensitivity analysis to the appendix.
- **Expand discussion on scope of gradient applicability.** We added discussion on scope of gradient applicability in *Learning Losses* and *Learning Algorithms*. In *Learning Losses* we now state directly that Theorem 6 is conditional on differentiability and tractability of the relevant optimal value-function and symbolic potential. Moreover, in *Learning Algorithms* we explain that the bilevel method works with first-order value-function gradients and clarify when modular learning or stochastic policy optimization are better alternatives.
- **Policy-gradient scalability discussion.** These timeouts are expected in more complex settings: IndeCateR sums over values of individual categorical decisions while sampling the remaining ones, which is more efficient than full joint enumeration but still leads to high-variance updates and requires many rounds of inference before a parameter update. Slower wall-clock convergence and timeouts are not surprising in these structured prediction settings, and variance reduction and more efficient policy-gradient estimators are promising directions for future work.

We also incorporated the minor clarifications requested by the reviewer, including theorem-numbering consistency and wording that avoids conflating baseline digit accuracy with consistency for Visual-Sudoku.

Reviewer 3

We appreciate the reviewer's thoughtful and constructive assessment, especially the recognition that the paper's central contribution and thesis are valuable and theoretically well grounded, that the DSVar/DSPar/DSPot tax-

onomy is clean, that the learning algorithms are coherent and well characterized, that the bilevel value-function optimization is a non-trivial and important contribution, and that the limitations section is honest and comprehensive.

Major Points

- **Experimental comparisons against other NeSy systems or baselines.** We scoped the empirical positioning so that the results are presented as studying NeSy-EBMs as instantiated through NeuPSL in the selected settings. We now make this clear in the *Abstract*, *Introduction* (Section 1), *Empirical Analysis* (Section 6), and *Limitations* (Section 7).

- **Clarify the learning loss separability assumption.** Our intent is to study the general empirical risk minimization formulation for the class of NeSy-EBM losses developed in Section 5, not to claim that every NeSy framework must use the same per-sample objective structure. This is not an IID assumption on inference, nor does it remove within-sample dependencies: in NeuPSL, prediction and the value-based and minimizer-based losses still depend on joint inference over the full target variable vector. We now make this even more clear in the *NeSy-EBM Learning* and *Learning Losses* subsections of 5.

- **The DSVar indicator function.** DSVar is used when the neural output is coupled with a subset of the target variables. The role of the indicator is to formalize that these neural predictions are treated in practice as fixed evidence during symbolic inference. We now emphasize this practical interpretation in *Modeling Paradigms for NeSy* in Section 3, while keeping the formal definition in the appendix *Extended Modeling Paradigms*.

- **Adding weight coefficients to learning losses.** We revised *Learning Losses* so that the aggregate objective now explicitly includes scalar coefficients $\lambda_1, \lambda_2, \lambda_3$. This makes the theoretical presentation consistent with the implementation and the reported hyperparameters.

- **Proof of Theorem 6.** We appreciate the reviewer's careful eye on this point. In Deep HL-MRFs, for instance, the perturbation is to the neural output provided to the symbolic component, and this can change the induced feasible set through $\Omega(\mathbf{x}_{sy}, \mathbf{g}_{nn}(\mathbf{x}_{nn}, \mathbf{w}_{nn}))$. Our intent in Theorem 6 is not to claim differentiability, but to state a powerful general result: the gradient form for NeSy-EBM value-functions is consistent whenever the relevant derivatives exist. The proof already makes this conditionality explicit in the sentence stating the assumption: *when f and $\overline{V}$ are right and left hand differentiable, respectively.* The theorem is framed as a statement about gradient form when differentiability exists at the point, not as a claim that differentiability must hold in every NeSy-EBM. For NeuPSL specifically, we consider the parameterized feasible-set in Theorem 5 using the LCQP formulation of NeuPSL.

Additional Remarks

- **Clarify overlap between "learning under constraints" and "post-training" in related work.** The learning from constraints section describes the broader paradigm of incorporating domain knowledge as a learning loss during training, whereas post-training refers to applying such NeSy learning objectives specifically when adapting pre-trained or foundation models to downstream tasks. We have added a clarifying sentence in the post-training subsection to explicitly state this distinction. Furthermore, the reviewer notes that both settings typically rely on soft constraints that serve as learning signals rather than enforcing hard guarantees at inference (as in the constraint satisfaction section). This is typically true; however, constraint satisfaction can also be incorporated into learning, for example, by performing a constraint-satisfaction step at each gradient update to ensure that predictions satisfy hard constraints (as in NeuPSL).

- **Clarify modeling paradigms applicability statement.** In this sentence, approaches refer to specific NeSy systems or methods (e.g., Logic Tensor Networks), while categories refer to the modeling paradigms introduced in Section 3. Our intent was to note that while some systems can be described by a single modeling paradigm, some approaches span multiple paradigms or combine elements of them. We have revised the sentence to clarify this distinction.

- **Statistical significance and standard errors in RoadR results.** The RoadR result in Table 3 is reported on a single train/test split and we did not report standard deviations or a statistical-significance test for that experiment. We revised the *Datasets and Models* and *Constraint Satisfaction and Joint Reasoning* subsections of 6 to make this protocol explicit and to calibrate the discussion accordingly. Our intended claim in RoadR is that the NeuPSL-instantiated *DSPar* NeSy-EBM enforces the logical requirements, improving

constraint satisfaction consistency from 27.5% to 100%, while maintaining comparable object detection prediction performance to the DETR baseline, rather than claiming a statistically established F1 improvement.

- **Provide Conditions under which differentiability holds.** We agree that this additional detail is useful. However, a full characterization of differentiability conditions for general NeSy-EBMs is beyond the scope of this paper, and our intent is not to claim such a characterization in Theorem 6. For NeuPSL specifically, *A Smooth Formulation of Deep HL-MRF Inference* gives the concrete smooth regularized setting used to obtain the continuity and differentiability results.

- **Address standard deviation over all folds in Figure 6** Figure 6 reports convergence traces for a single fold, and we did not present standard deviations for that figure. We revised the *NeSy-EBM Learning* subsection of Section 5 to make this even clearer. Our intended claim is that Figure 6 provides a qualitative illustration of the convergence and wall-clock tradeoffs among the Energy, IndeCateR, and Bilevel NeSy-EBM learning algorithms on the reported fold.

- **LTNs as NeSyEBMs.** We agree that the full generality of LTNs is broader than the particular NeSy-EBM view presented here. We clarified in the appendix *Expressing NeSy Approaches via NeSy-EBMs*, at the beginning of the section, that all the systems we present are commonly used instantiations, that this representation limits the full expressiveness of each system, and that alternative semantics may require different symbolic energy constructions.

References

T. R. Besold, A. S. d'Avila Garcez, S. Bader, H. Bowman, P. M. Domingos, P. Hitzler, K. Kühnberger, L. C. Lamb, D. Lowd, P. M. V. Lima, L. de Penning, G. Pinkas, H. Poon, and G. Zaverucha. Neural-symbolic learning and reasoning: A survey and interpretation. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 2022.

L. De Raedt, S. Dumančić, R. Manhaeve, and G. Marra. From statistical relational to neuro-symbolic artificial intelligence. In *IJCAI*, 2020.

C. Dickens. *A Unifying Mathematical Framework for Neural-Symbolic Systems*. PhD thesis, University of California, Santa Cruz, Santa Cruz, CA, USA, 2024.

C. Dickens, C. Gao, C. Pryor, S. Wright, and L. Getoor. Convex and bilevel optimization for neuro-symbolic inference and learning. In *ICML*, 2024a.

C. Dickens, C. Pryor, and L. Getoor. Modeling patterns for neural-symbolic reasoning using energy-based models. In *AAAI Spring Symposium on Empowering Machine Learning and Large Language Models with Domain and Commonsense Knowledge*, 2024b.

E. Giunchiglia, M. C. Stoian, and T. Lukasiewicz. Deep learning with logical constraints. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

C. Pryor. *Foundations of Neural-Symbolic AI: Architecture and Design*. PhD thesis, University of California, Santa Cruz, 2024.

C. Pryor, C. Dickens, E. Augustine, A. Albalak, W. Y. Wang, and L. Getoor. NeuPSL: Neural Probabilistic Soft Logic. In *IJCAI*, 2023.

E. van Krieken, E. Acar, and F. van Harmelen. Analyzing differentiable fuzzy logic operators. *Artificial Intelligence (AI)*, 302:103602, 2022.