

Authors' Response to Reviewers

Neurosymbolic Inference, Reasoning and Proving in Propositional Logic: A Survey

Nenad Krdzavac and Felix Engel

Tracking number of the previous submission #: 931-1953

Responsible editor: Artur Garcez

Response to the Editor and Reviewers

Dear Editor and Reviewers,

We sincerely thank the editor and both reviewers for their careful, constructive, and encouraging evaluation of our manuscript. We have considered every comment and substantially revised the manuscript.

For clarity, each reviewer comment is reproduced below, followed by our response, a description of the changes made, an excerpt from the revised manuscript where appropriate, and the corresponding section, page, and line numbers.

The marked manuscript uses the following colour scheme:

- **Blue:** revisions addressing Reviewer 1;
- **Green:** revisions addressing Reviewer 2;
- **Violet:** revisions addressing comments raised by both reviewers.

Page and line references in this response refer to the clean revised manuscript. The clean and marked manuscripts contain identical text and line numbering.

Summary of the Major Revisions

The principal revisions are summarised as follows:

1. We clarified the definitions, organization, and interpretation of the X–Y–Z taxonomy.
2. We distinguished the taxonomy hierarchy from the capability and assurance evaluation scales.
3. We clarified the literature-search and paper-selection methodology.
4. We revised the historical-period analysis and the presentation of temporal trends.
5. We reorganized and expanded the introduction, background, survey methodology, and conclusions.
6. We corrected terminology, references, tables, grammatical errors, and typographical issues throughout the manuscript.

Response to Reviewer 1

We thank Reviewer 1 for recognizing the relevance and novelty of the proposed taxonomy and for providing detailed recommendations that helped us improve the paper’s conceptual clarity and presentation. In the responses below, we explain how each comment has been addressed and indicate the corresponding locations in the revised manuscript. The final PDF presents a clean version of the revised manuscript, with deleted or replaced passages removed. The marked version shows how the text was revised and how the reviewer’s comments were addressed. Changes in text are highlighted in Blue and Violet.

Reviewer 1 – Comment R1.1

I think a taxonomy such as the one being proposed is certainly needed, and the paper has opened my eyes to how other recent works (including my own!) overlook the issue of reasoning and proving capabilities and assurance. I also give kudos for drawing attention to older, more traditional NeSy works, as I believe there is a misconception that neurosymbolic computing is a relatively new field. So in terms of relevance and novelty this survey hits the mark.

Authors’ response: We thank the reviewer for this encouraging assessment. We appreciate the recognition of the taxonomy’s relevance and novelty, particularly its emphasis on reasoning and proving capabilities, assurance, and foundational NeSy research.

Reviewer 1 – Comment R1.2

However, I’m afraid this paper seems some major revisions before I feel it is ready for publication, but I am enthusiastic in encouraging such revisions. My primary concern is a lack of clarity of the proposed taxonomy, especially with respect to interpreting the ordering and hierarchy. Secondary concerns would be a need to expand on how "eras" are defined, a lack of summarised findings and future work proposals in the concluding chapter, a bit of uncertainty with respect to target audience, and a large collection of typographical issues.

Authors’ response: We thank the reviewer for the constructive comments and encouragement to revise the paper. We revised the X–Y–Z taxonomy and the **A Taxonomy for NeSy Frameworks Survey** section according to the suggestions of both reviewers, with particular attention to clarifying the ordering and hierarchy. A detailed description of these revisions is provided in our response to comments [R2.7](#), [R2.8](#) and [R2.10](#).

Text in the marked manuscript

Location in the marked manuscript: Section 3, pages 16–17, lines 453–493; Section 3.1, pages 17–18, lines 495–510

Regarding the definition of **eras**, we replaced them with publication periods grouped by decade, from 1940–1949 through 2020–2026. The final period covers the years available up to 2026. Further details of the literature search conducted for each period are provided in our response to comment [R1.36](#).

Text in the marked manuscript

Location in the marked manuscript: Section 5, page 46, lines 1265–1288.

We also corrected typographical and grammatical errors throughout the manuscript, with support from [LanguageTool](#), [DeepL](#), and [OpenAI](#) tools.

In response to the concern regarding the Conclusion, we have revised the Conclusion to summarize the main findings, explain their implications for assessing NeSy research, and identify concrete directions for future work. In particular, we emphasize the imbalance between reported PL capabilities and documented assurance mechanisms, while clarifying the limits of the annotation-based evaluation. The revised section also discusses proof-producing methods, independent validation, improved reporting and annotation procedures, and extensions to other logics. Please also see our responses to Comments [R1.47](#) and [R1.48](#).

Text in the marked manuscript

Location in the marked manuscript: Section 7, pages 100-102, lines 3018–3089

Reviewer 1 – Comment R1.3

(1) Suitability as introductory text, targeted at researchers, PhD students, or practitioners, to get started on the covered topic.

To me it wasn't clear what level of background knowledge the reader is assumed to have. The authors give a very detailed (and good!) introduction to propositional logic, and skim over neural networks. Nothing wrong with this in itself. Based on this I would assume this paper was written to somebody very familiar with the latter and very unfamiliar with the former, but this is not clarified.

Authors' response: Thank you for this observation. We added a short paragraph at the beginning of the **Background** section clarifying the intended readership and assumed level of prior knowledge.

Text in the marked manuscript

Location in the marked manuscript: Section 2, lines 158–163.

Reviewer 1 – Comment R1.4

Furthermore, assuming that somebody completely new to propositional logic reads section 2, I get the feeling that some of the terminology introduced in chapter 3 is a bit of a jump. For example terms like "lemma" and "witness" are used in chapter 3 but not defined here or in section 2. Note also my below comment about distinguishing between general and specific cases in the hierarchy - for a novice in propositional logic this may be difficult to appreciate.

Authors' response: To make the terminology accessible to readers new to propositional logic, we now define **lemma** and **witness** at the end of Section 2.3. We also explain **clause-based proof**, **resolution-based proof** in Section 2.4. These additions introduce the terms before they are used in the taxonomy descriptions in Section 3.

Text in the marked manuscript

Location in the marked manuscript: Section 2.3, page 12, lines 303–309; Section 2.4, pages 12-13, lines 311–353.

Reviewer 1 – Comment R1.5

(2) How comprehensive and how balanced is the presentation and coverage
A wide range of other reviews are covered, with 43 specific methods are selected in the survey. While not a deal-breaker for me, I couldn't help but feel this felt disproportionately small considering how many methods must be covered by the other reviews overall. Either more papers should be analyses, or a rationale for selection should be clearly specified.

Authors' response: We revised the literature search methodology and expanded the survey from 43 to 181 publications in total. A detailed explanation of the revised search methodology and the criteria used to select publications is provided in our response to comment [R1.36](#).

Text in the makred manuscript

Location in the marked manuscript: Section 5, page 46, lines 1265–1288.

Reviewer 1 – Comment R1.6

As a positive, papers spanning many years are covered, which is very nice to see considering people sometimes misconceive that neurosymbolic is a new field.

Authors' response: By covering studies across several decades, we aim to highlight the historical development of NeSy research and clarify that its foundations predate recent advances.

Reviewer 1 – Comment R1.7

(3) Readability and clarity of the presentation.
Though I nonetheless found the subject matter to very interesting and relevant, I'm afraid to say clarity is my major concern.
Firstly, perhaps some of my confusion would be relieved if I was familiar with the SKOS system, but it is new to me. Authors may need to elaborate on this in their preliminaries if its fundamental to understanding the taxonomy, or clarify that they assume the reader to be familiar with it (see above point about target audience).

Authors' response: To improve clarity, we have added a short explanation of ontologies and SKOS in Section 2.7. This section introduces the basic terminology and explains how SKOS represents concepts and their broader and narrower relationships. It provides the background needed to understand the taxonomy presented in Section 3 without assuming prior familiarity with SKOS.

Text in the makred manuscript

Location in the marked manuscript: Section 2.7

Reviewer 1 – Comment R1.8

In chapter 3 authors discuss the notion of "levels" of capability/assurance. At first when reading chapter 3 I thought "levels" referred to hierarchy depth (e.g. X1.1 is higher level than X1.1.1.) but I didn't appreciate until section 6 that "level" applies to the "a" in "1.a". Section 6, second paragraph: "A higher value of a indicates a higher reasoning capability level". It would have been helpful to have this explained in chapter 3. Anyway, assuming then this means that, for example, X1.8 has higher reasoning capability than X1.1., I'm not really clear on how this aligns with the titles assigned to the various levels.

For example, what does it mean for a "robust and interpretable" (X1.8) solution to have higher reasoning capability than a solution that offers "search and control" (X1.4)? This suggests "All Robustness and Interpretability methods enable search and control but not all search and control methods enable robustness and interpretability". Are the authors saying that it is impossible for a model to be robust and interpretable without search and control functionality?

Authors' response: We thank the reviewer for identifying this ambiguity. We agree that the original terminology and the statement "A higher value of a indicates a higher reasoning capability level" **incorrectly suggested an ordering of capabilities**. We have replaced this interpretation in Sections 3 and 6. In the revised taxonomy, numerical identifiers distinguish concepts; they do not denote ordinal levels of capability or assurance.

Text in the marked manuscript

Location in the marked manuscript: Section 3, page 17, lines 476–480

Our SKOS-based taxonomy uses `skos:broader` and `skos:narrower` relations to connect more general and more specific concepts. Hierarchy depth indicates conceptual specificity, rather than a higher or lower capability or assurance level. These relations do not establish capability rankings or logical implications between the properties of annotated frameworks. To make this distinction explicit when the taxonomy is introduced, we have added the following explanation to Section 3: **They do not indicate rankings of capability, sophistication, or performance, or progression between categories. A child concept is more specific than its parent.**

Text in the marked manuscript

Location in the marked manuscript: Section 3, page 17, lines 476–480

Consequently, X8: reliability is not assumed to represent a higher reasoning capability than X1: Logical foundation or X4: Search strategy. These categories describe different characteristics. We do not claim that all robust and interpretable frameworks provide search and control, or that robustness and interpretability require such functionality. A framework may exhibit either characteristic or both. Surveyed papers are annotated with the corresponding concepts according to the evidence they provide about the frameworks they describe.

We have also revised the evaluation in Section 6, as formally defined in Section 6.1. The numerical components of taxonomy identifiers are no longer used to determine evaluation scores. Instead, the Z scores are computed from corpus-based concept frequencies and annotation coverage. Any score difference between papers annotated with X8 and X4 therefore results from the defined evaluation procedure and their assigned concept sets, rather than the numerical order of these identifiers. Such a difference does not establish an inherent ranking between the concepts or imply that one characteristic requires the other.

Text in the marked manuscript

Location in the marked manuscript: Section 6.1, page 89-92, lines 2766–2814

Reviewer 1 – Comment R1.9

I think it's great that the authors are providing a means to categorise NeSy work according to terms such as robustness, interpretability, search, control, etc., and according to propositional logic more generally, but I'm afraid I just don't understand the ordering of levels.

Authors' response: We thank the reviewer for highlighting this lack of clarity. We revised the X–Y–Z taxonomy and clarified the interpretation of its hierarchy. Within the X and Y hierarchies, numerical identifiers distinguish categories; they do not indicate a ranking, increasing capability, or a progression between categories. A child concept represents a more specific concept within its parent category. The Z dimension provides **evaluation metric**, as explained separately in the revised manuscript.

We revised the **A Taxonomy for Surveying NeSy Frameworks** section to make these distinctions explicit. A detailed explanation of the taxonomy revisions is provided in our response to comment [R2.7](#), [R2.8](#), and [R2.10](#).

Text in the marked manuscript

Location in the marked manuscript: Section 3, page 16, starting line 452

Reviewer 1 – Comment R1.10

I'm also a bit unclear on cases where a category of the hierarchy only breaks down into a single sub-group. For example the only child of X1.1.1 Decision Mechanisms is X1.1.1.1."Basic decisions", which implies the existence of "Complex decisions" or "Advanced Decision", which the hierarchy or accompanying text don't address. And again there is X1.1.1.1.1 "Decision-only", suggesting a "Decision-not-only". Maybe 1.1.1.1 and 1.1.1.1.1 are special cases, but the accompanying text doesn't elaborate on the definition of the general case. Any reader new enough to propositional logic to need to read section 2 is going to need help filling in these gaps (also relates to my later comments on target audience).

Authors' response:

We thank the reviewer for highlighting this ambiguity. We revised the X–Y–Z taxonomy and merged categories containing only one subgroup with their parent concepts where possible. These revisions clarify the hierarchy and remove unnecessary distinctions that could imply additional, undefined categories. Detailed explanations are provided in our responses to comments [R2.7](#), [R2.8](#), and [R2.10](#).

Text in the marked manuscript

Location in the marked manuscript: Section 3.

Reviewer 1 – Comment R1.11

Some elements of X and Y have counterparts (e.g. X1.7.1 and Y1.1.1). How are they distinct and/or what is the relationship between them? More clear definitions of capability (X) and assurance (Y) and their relationship would help here.

Authors' response: We thank the reviewer for highlighting the need to clarify this distinction. We revised the X–Y–Z taxonomy and added definitions of capability (X) and assurance (Y)

at the beginning of the Section 3: A Taxonomy for Surveying NeSy Frameworks section. *X* describes the reasoning-related functions or characteristics exhibited by a framework, whereas *Y* describes the mechanisms or properties that support confidence in its reasoning or results. Related concepts in these dimensions capture complementary aspects and are assigned according to the evidence reported in the publication. An *X* annotation does not automatically imply a corresponding *Y* annotation. Further details of the taxonomy revisions are provided in our response to comments [R2.7](#), [R2.8](#), and [R2.10](#).

Text in the marked manuscript

Location in the marked manuscript: Section 3

Reviewer 1 – Comment R1.12

Regarding *Z*, as I read sections 4 and 5 it wasn't clear to me why *Z* wasn't included in the analyses of other literature reviews (section 4) and NeSy methods (section 5). I looked at tables 10-16 and wondered where *Z* was. It only made sense when I started reading chapter 6. Authors should please: - Clarify in section 3 that *Z* is not relevant to sections 4 and 5, and why it is not relevant. - Elaborate on *Z* more in section 3 (possibly moving all explanation of *Z* methodology here or at least a summary) - at the very least please explain that it is derived from *X* and *Y*. - Clarify what they mean by "evaluation". In a sense analyses in sections 4 and 5 are also evaluations, just qualitative ones, whereas section 6 is a quantitative evaluation (on data gathered in section 5).

Authors' response: We thank the reviewer for identifying the need to clarify the role of the *Z* dimension before presenting the analyses in Sections 4 and 5. We have revised Subsection 3.4 *Z: Evaluation Metric for NeSy Frameworks*, to address each point as follows.

1. **Why *Z* scores are not included in Sections 4 and 5?** In Subsection 3.4, we now explain that Section 4 qualitatively compares existing literature reviews and their methodologies and taxonomies, rather than evaluating individual NeSy frameworks using framework-level *X* and *Y* annotations. Section 5 establishes and explains the *X* and *Y* annotations of the surveyed frameworks. The *Z* scores are computed subsequently in Section 6, because their computation requires corpus-wide concept frequencies and normalization values. Thus, their absence from the annotation tables reflects the separation between qualitative annotation and subsequent quantitative evaluation.

Text in the marked manuscript

Location in the marked manuscript: Section 3.4, pages 29–30, lines 740–772.

2. **Summary of the *Z* methodology and its derivation from *X* and *Y*.** We have expanded Subsection 3.4 to explain explicitly that *Z* scores are derived numerical measures, rather than concepts assigned directly during annotation. The subsection now summarizes concept-frequency weighting, raw-score normalization, normalized coverage, and scaling to the interval $[0, 6]$. It explains that *Z*1 and *Z*2 combine the normalized frequency-weighted scores with coverage in the *X* and *Y* dimensions, respectively. It also clarifies that *Z*3 is the scaled harmonic mean of the normalized *X*-coverage and *Y*-coverage values, rather than of *Z*1 and *Z*2. The detailed mathematical definitions remain in Section 6.1, which is explicitly referenced from Subsection 3.4.

Text in the marked manuscript

Location in the marked manuscript: Section 3.4, pages 29-30, lines 754–762.

3. **Meaning of “evaluation”.** We agree that the analyses in Sections 4 and 5 also constitute evaluation. Subsection 3.4 now states that evaluation in this survey includes both qualitative analysis and quantitative assessment. It distinguishes the qualitative comparison of existing literature reviews in Section 4, the qualitative analysis and annotation of NeSy frameworks in Section 5, and the quantitative evaluation in Section 6. We explicitly state that the annotations established in Section 5 supply the data used to compute the Z scores.

Text in the marked manuscript

Location in the marked manuscript: Section 3.4, page 29, lines 740–753

Reviewer 1 – Comment R1.13

Z1.3: If the process is supposed to be automated (page 22, last paragraph) how does one automatically decide on the choice of 1, 2, 3 for A in equation 13? I’m also not clear how what it means for integration of neural learning and formal inference to be “strong” or “weak”. These points unclear I unfortunately found it difficult to appreciate what any of the discussions on Z1.3 were saying.

Authors’ response: We thank the reviewer for identifying this ambiguity. We agree that the original choice of $A \in \{1, 2, 3\}$ and the descriptions of neurosymbolic integration as *strong* or *weak* were insufficiently defined. In the revised evaluation, we have removed this parameter and replaced the former integration measure with **Z3: Joint coverage** concept. The revised methodology therefore does not require selecting a value for A or judging integration strength.

Section 6.1 now formally defines the computation of the three scores as follows.

1. **Capability level ($Z1$).** Each X concept receives a weight equal to the proportion of surveyed papers annotated with that concept. For each paper, these weights are summed and divided by the maximum raw capability score observed in the corpus (180 in this case). The result is multiplied by the paper’s normalized X -coverage and scaled by six. Normalized X -coverage is the number of distinct X concepts assigned to the paper divided by the maximum number assigned to any single paper in the surveyed corpus.
2. **Assurance level ($Z2$).** This score is computed analogously using the assigned Y concepts, their corpus-wide frequencies, and the paper’s normalized Y -coverage. Thus, both $Z1$ and $Z2$ reflect concept prevalence and annotation breadth, rather than the numerical structure of taxonomy identifiers.
3. **Joint coverage ($Z3$).** This score is the harmonic mean of the normalized X -coverage and Y -coverage values, scaled by six. It is computed from these coverage values, rather than from $Z1$ and $Z2$, and does not incorporate concept-frequency weights. A high score requires substantial coverage in both dimensions. It is zero when either coverage value is zero and reaches six only when both coverage values equal one.

We also clarify the scope of automation. Assigning X and Y concepts to publications requires an evidence-based qualitative assessment. Once these annotations are established, concept frequencies, weights, normalization values, and the three scores can be computed automatically

using the definitions in Section 6.1. Automation therefore concerns the numerical computation from the annotations, rather than an automatic judgement of integration strength.

The revised scores provide corpus-relative summaries of the annotated capability and assurance profiles. In particular, *Z3* measures joint coverage and does not establish the strength of neurosymbolic integration, task performance.

Text in the marked manuscript

Location in the marked manuscript: Section 6.1.

Reviewer 1 – Comment R1.14

I think some improved notation and formatting would also help the taxonomy be more accessible. - For example, the first digit in any category's enumeration is redundant: X1.a could be X.a and Y1.a could be Y.a with no information loss. With the redundant 1 made the hierarchy feel 1 level deeper than it was, which was a bit intimidating when already trying to navigate a hierarchy of depth 4. - Breaking down sections 3.1 and 3.2 with further sub-subsections at least for each [X/Y].a would make the explanation of the hierarchy a bit easier to follow. For example a 3.1.1 sub-subsection for "foundational reasoning".

Authors' response: We thank the reviewer for this helpful suggestion. We revised the notation and formatting of the X–Y–Z taxonomy to make the hierarchy easier to read and less visually intimidating. In particular, the redundant first digit was removed from the taxonomy identifiers. Thus, identifiers such as X1.2, Y1.2, and Z1.2 were replaced by the shorter forms X2, Y2, and Z2, respectively. This change preserves the same taxonomic information while reducing the apparent hierarchy depth.

We also restructured the explanatory sections. The former Section 3.1 is now Section 3.2, and the former Section 3.2 is now Section 3.3, and Section 3.3 is now Section 3.4. Sections 3.2 and 3.3 were broken into clearer subsections using the first-level taxonomy concepts as subsection titles. For example, the explanation of the X dimension is now organized around subsections corresponding to the main X concepts, and the explanation of the Y dimension is organized analogously around the main Y concepts. This makes the taxonomy easier to navigate and allows readers to connect each explanatory paragraph directly to the corresponding taxonomy branch.

Text in the marked manuscript

Location in the marked manuscript: Section 3.2, Section 3.3.

Reviewer 1 – Comment R1.15

(4) Importance of the covered material to the broader Neurosymbolic AI community. The paper is very important in this respect. I feel that the issues of validation, proof etc. have been overlooked in recent years, at least with respect to what I've read. There's been disproportionate attention given to explainability and interpretability compared to other properties of neurosymbolic models. This paper was a wake-up call to me at least! This is the main reason I want to re-emphasise that I do strongly encourage the authors to address my concerns, because I would very much like to see this published.

Authors' response: We thank the reviewer for these encouraging comments and for recog-

nising the importance of this survey to the broader Neurosymbolic AI community. We share the reviewer’s view that validation, soundness, and proof deserve greater attention alongside explainability and interpretability. We particularly appreciate the reviewer’s support for publication and constructive criticism, which have guided our revisions. We have carefully considered the concerns raised, revised the paper accordingly, and provided the locations of the revised text in our responses.

Reviewer 1 – Comment R1.16

Eras:

In the background review, I was not clear on the rationale used to partitioned years into eras. For each era could use some justification for the choice of name and perhaps more importantly why one era ends and the next begins. This is important given that they influence the plots in chapter 6, which would appear differently to somebody who may decide to partition the years differently.

Authors’ response: To avoid introducing subjective boundaries based on selected methodological developments, the surveyed papers are grouped into fixed nine periods, with publication year as the sole criterion for assigning a paper to a period. Period labels are therefore chronological rather than methodological, while the dominant technical developments are discussed separately. The chronological coverage includes the 1940s, 1950s, 1960s, 1970s, 1980s, 1990s, 2000s, 2010s, and 2020s. The 1943 contribution [McCulloch and Pitts \(1943\)](#) is retained as a foundational antecedent but is not considered in the statistical plots (see Section 6, page 89, lines 2742-2747). Its taxonomy annotations and evaluation values are reported for completeness, but it is excluded from the period-level aggregate statistics and trend plots. The continuous body of eligible NeSy work in the surveyed corpus begins in 1990. Finally, the 2020–2026 period is incomplete and is identified as such when publication counts are interpreted.

Location in the marked manuscript: Section 5

Reviewer 1 – Comment R1.17

That said, the plots in chapter 6 may be better plotted by calculating Z values for each actual year of publication. Given that authors have the years the papers were published they easily have the data available to plot at this level of granularity along the horizontal axis. It will give a more accurate picture of how the Z dimensions change over time that is not impacted by how they or anybody else chooses to partition the different eras and I’d be very interested to see this. They could still plot the eras as annotations over the more fine-grained plot.

Authors’ response: We agree that plotting the evaluation results by actual publication year provides a more detailed account of temporal variation than aggregation into chronological periods alone. We have revised Section 6, particularly Subsection 6.5, to implement this suggestion and explain the resulting patterns.

The revised evaluation first computes $F(p) = (Z_1(p), Z_2(p), Z_3(p))$ for each of the 180 evaluated publications, using the definitions in Subsection 6.1. Here, Z_1 describes the annotation-based capability profile, Z_2 describes the assurance profile, and Z_3 describes joint coverage of the X and Y dimensions. The publication-level scores are then grouped by actual publication year to calculate yearly means, medians, and binned Shannon entropy separately for each metric. Concept frequencies, weights, and normalization factors are calculated across the evaluated corpus, rather than separately within each year, so the yearly statistics use a common scoring basis. Figure 8 in Section 6 now explicitly distinguishes the calculation of publication-level scores

in Step 4 from yearly aggregation and plotting in Step 5.

We have added three figures presenting these complementary summaries: In Section 6, Figure 9 shows yearly arithmetic means, Figure 10 shows yearly medians, and Figure 11 shows yearly binned Shannon entropy. Each figure contains separate panels for Z_1 , Z_2 , and Z_3 , with actual publication years along the horizontal axis. Following the reviewer’s suggestion, the chronological periods G5–G9 are retained as visual annotations; their boundaries do not determine the yearly aggregation. Years without evaluated publications appear as gaps in the statistical plots rather than being assigned zero scores. The historical reference [McCulloch and Pitts \(1943\)](#) remains in the survey for context but is excluded from the quantitative evaluation, as explained at the beginning of Section 6 and Subsection 6.2.

We have also expanded the accompanying explanation to clarify what each statistic contributes. The mean describes the average score in a given year, while the median describes its central value and is less sensitive to extreme scores.

To support interpretation, each figure includes a bottom panel showing the number of evaluated publications per year. This addition makes clear when an apparent peak is based on very few publications. For example, the highest yearly mean and median of Z_1 , both 3.896 in 2007, are based on a single publication. Its empirical entropy is consequently zero. In contrast, the highest observed yearly entropy of Z_1 , 2.002 bits in 2024, is based on 18 publications. The revised discussion therefore considers publication counts alongside the statistics and avoids interpreting isolated peaks as evidence of sustained progress.

The yearly plots also reveal variation that would be obscured by a single summary for each chronological period. For example, within G9, the mean of Z_1 rises to 1.897 in 2024, falls to 0.767 in 2025, and rises again to 1.119 in 2026. Furthermore, the yearly means of Z_2 and Z_3 are positive in several years for which their medians remain zero. The revised text explains how these differences distinguish the presence of nonzero scores in a minority of publications from the central pattern within a year. These observations describe the evaluated corpus and do not establish statistically significant temporal trends or directly measure framework quality.

Finally, the computations and figure generation are implemented in the `nesy_evaluation` Python project, referenced in Listing 1 of Appendix A.1. All yearly statistics are available in `data/processed/z_statistics_by_year.csv`. The revised subsection concludes with guidance for readers, particularly PhD students, on distinguishing evidence of reasoning capability from evidence of assurance and on consulting the underlying concept annotations alongside the numerical summaries.

Text in the marked manuscript

Location in the marked manuscript: Section 6.

Reviewer 1 – Comment R1.18

Typographical issues:

There are many spelling and grammatical errors throughout. I don’t wish to list them all since they may be eliminated once the more major corrections are addressed anyway. But authors should please run a grammar and spell-check before resubmitting.

Authors’ response: We thank the reviewer for this comment. We checked the manuscript for spelling and grammatical errors using an OpenAI tool, DeepL, and LanguageTool (<https://languagetool.org/>). We manually reviewed the suggestions and revised the text accordingly. We applied the same approach to the other sections. The resulting grammatical revisions are

marked in **Violet** in the revised manuscript.

Reviewer 1 – Comment R1.19

Introduction

Authors should maybe elaborate more on why these guarantees, proving capabilities and assurance levels are important. What are some real-world application areas where this could have critical consequences? This sort of opening discussion helps motivate the reader to keep reading. If people are new to formal systems, what motivates them to learn about it and frame theirs or other NeSy methods in this way?

It's very relevant to recent discussions on guaranteed safe ai, actually (<https://arxiv.org/pdf/2405.06624?>), where formal guarantees of AI behaviour are encouraged.

Authors' response: We thank the reviewer for this helpful comment and for suggesting the reference on Guaranteed Safe AI. We have expanded the Introduction to explain why reasoning and proving capabilities and assurance mechanisms matter, particularly in applications such as autonomous transport, medical devices, and industrial control, where incorrect decisions can cause injury or damage.

The added paragraph connects this practical motivation to the scope of our survey. It explains how satisfying assignments and checked UNSAT refutations provide different forms of evidence for problems represented in PL, and distinguishes generating a proof from validating it. This gives readers unfamiliar with formal systems a concrete reason to understand these concepts and use them when assessing NeSy frameworks: empirical performance alone does not establish satisfaction of a specified requirement.

We also connect this discussion to the Guaranteed Safe AI perspective proposed by "[davidad Dalrymple et al. \(2024\)](#)", while clarifying that formal guarantees depend on the model, specification, and verification assumptions. Finally, we explain how distinguishing capabilities from assurance properties helps readers compare frameworks, assess the evidence supporting their results, and identify limitations relevant to their applications.

Text in the marked manuscript

Location in the marked manuscript: Section 1

Reviewer 1 – Comment R1.20

Please provide a reference to the SKOS ontology in paragraph 2

Authors' response: The SKOS ontology is first cited in Section 1 and subsequently in Section 2.7, which also provides a URL to its raw file.

Text in the marked manuscript

Location in the marked manuscript: Section 1, page 6, lines 112-116; Section 2.7

Reviewer 1 – Comment R1.21

Maybe also elaborate more on the scope. Why stop at propositional logics? What about predicate logic? Temporal logic? For example.

Authors' response: We focus on PL to establish a clearly defined scope within which NeSy

reasoning and proving capabilities, together with their assurance mechanisms, can be compared systematically. PL already encompasses substantial research on satisfiability, logical consequence, and proof generation and validation. Extending the survey to predicate or temporal logic would introduce additional distinctions that the current taxonomy does not explicitly represent, including quantification, variable binding, and reasoning in predicate logic, and temporal operators and reasoning over states and execution traces in temporal logic. Integrating these features with neural networks also introduces different representation and inference requirements.

Consequently, covering these logics would require a substantial extension of the taxonomy and potentially separate, interconnected taxonomies. A more expressive ontology could support this extension through explicit class hierarchies and axioms describing relationships and constraints among capabilities, logical formalisms, and assurance mechanisms. Complementary SHACL shapes could validate the structure and consistency of framework annotations against specified requirements. We therefore regard extending the classification to other logics as an important direction for future work, beyond the scope of the present survey.

Text in the marked manuscript

Location in the marked manuscript: Section 7, pages 101–102, lines 3078–3089.
Section 1.1, page 5, lines 70–79.

Reviewer 1 – Comment R1.22

Background

Positions of 2.6 and 2.5 should perhaps be swapped. At present 2.5. is introducing neuro-symbolic before neural networks are introduced.

Authors' response: Sections 2.5 and 2.6 are swapped.

Reviewer 1 – Comment R1.23

2.1. Propositional Logic: Just want to reemphasise I think this is a good introduction to propositional logic.

Authors' response: We sincerely thank the reviewer for this positive and encouraging comment. We are pleased that the introduction to propositional logic in Section 2.1 was found clear and useful.

Reviewer 1 – Comment R1.24

2.2. Inference in PL: I'm not sure if both the last paragraph of this section or the first paragraph of 2.3. are necessary as they both say roughly the same thing. I'd lean towards removing the last paragraph of 2.2.

Authors' response: We thank the reviewer for this suggestion. We have deleted the last paragraph of Section 2.2 to avoid repetition with the first paragraph of Section 2.3

Location in the marked manuscript: Section 2.2.

Reviewer 1 – Comment R1.25

2.6. Short intro to ANNs: Maybe elaborate here that these form the basis of modern AI such as LLMs, CNNs, GNNs, etc.

Authors’ response: We thank the reviewer for this suggestion. We have expanded previous Section 2.6 (now it is Section 2.5) to explain that ANNs underpin many modern AI systems and to introduce convolutional neural networks (CNNs) and graph neural networks (GNNs), with supporting references. This addition connects the basic ANN concepts to these architectures while keeping the background section focused on the needs of the survey.

Text in the marked manuscript

Location in the marked manuscript: Last paragraph in Section 2.5, page 14, lines 365–371

Reviewer 1 – Comment R1.26

3. Taxonomy for NeSy Frameworks Survey

Note my above suggestion of using subheadings that reflect the taxonomy, at least at the 1.a level. It might help to make this section easier to digest. It was tricky to get a feel for where I was in the hierarchy as I read through the paragraphs. The redundant 1 after [X/Y/Z] make this a bit more confusing too.

Authors’ response: We have added subsections within Section 3.2, “X: NeSy Reasoning and Proving Capability Level in PL,” and Section 3.3, “Y: NeSy Reasoning and Proving Assurance Level in PL,” with headings that reflect the corresponding taxonomy categories. These headings make the hierarchical structure more explicit and help readers locate each discussion within the taxonomy

Text in the marked manuscript

Location in the marked manuscript: Section 3.2, Section 3.3.

Reviewer 1 – Comment R1.27

Please introduce the SKOS ontology a bit of an introduction. Why was this chosen as opposed to other taxonomies?

Authors’ response: We thank the reviewer for this helpful suggestion. We have added Section 2.7, “SKOS: A Simple Knowledge Organization System,” which introduces ontologies and the basic SKOS concepts and properties, with appropriate references. We also clarify that SKOS is a standardized model for representing taxonomies, rather than an alternative domain taxonomy. We selected SKOS because its W3C-standardized vocabulary supports the labels, definitions, identifiers, and hierarchical relationships required by our classification, while facilitating sharing and reuse. Our purpose is to organize concepts for annotating and comparing NeSy frameworks. It does not require representing each category as an OWL class with logical restrictions.

Text in the marked manuscript

Location in the marked manuscript: Section 2.7, page 16, lines 446–451.

Reviewer 1 – Comment R1.28

3.1. X: Capability: Page 6, first column, last sentence of penultimate paragraph: please elaborate on global versus local heuristics.

Authors’ response: We thank the reviewer for this suggestion. We have clarified the distinction between global exploration and local improvement, using single-variable flips in GSAT as an example of local search. We also clarify that these strategies can be combined.

Text in the marked manuscript

Location in the marked manuscript: Section 3.2.4, pages 22-23, lines 601–609.

Reviewer 1 – Comment R1.29

3.2. Y: Assurance
What is an assurance level?

Authors’ response: We have revised the introductory text of the Y-dimension section (Section 3.3) to define assurance level as the documented properties and mechanisms that support confidence in a NeSy framework’s PL task results. We provide examples, including soundness guarantees, proof validation, and certificate support, and clarify that Y concepts are assigned based on evidence reported in the publication. The term does not denote an ordinal ranking or require a framework to exhibit all characteristics in the Y dimension.

Text in the marked manuscript

Location in the marked manuscript: Section 3.3, pages 25–26, lines 649–656.

Reviewer 1 – Comment R1.30

"Interactive explainability (Y1.2.1) is a narrower concept of Explainability for assurance (Y1.2) in the X–Y–Z taxonomy, which generates step-by-step, human-interpretable reasoning traces." - I agree interactive vs non-interactive explanation is an important distinction with significant impact on user experience, but why does explainability have to be interactive to be human-interpretable? Why can’t non-interactive explanations be human-interpretable?

Authors’ response: We agree that human interpretability does not require interactivity. A non-interactive explanation, such as a fixed proof trace or a step-by-step derivation, can also be human-interpretable. Interactivity instead concerns whether users can engage with an explanation, for example by requesting clarification or exploring particular reasoning steps. It does not, by itself, guarantee interpretability or correctness.

In the revised Y dimension, we have replaced the previous explainability hierarchy with a single concept, Y2: **Interactive explanation**, described in Section 3.3.2. This concept specifically captures support for user interaction with explanations. It does not imply that non-interactive explanations cannot be human-interpretable or contribute to assurance.

However, Table 13 specifies evidence requirements, exclusions, and borderline decisions for coding interactive explanation (Y2), while Tables 3–12 in Section 3 provide corresponding coding guidance for other concepts in the X and Y dimensions.

Text in the marked manuscript

Location in the marked manuscript: Section 3.3.2, page 28, lines 708–724. Page 29, lines 733–738.

Reviewer 1 – Comment R1.31

Note that by "interactive" I'm imagining a system where a user's own line of enquiry dictates which explanations are displayed, for example the option to expand or collapse an explanation depending on the user's interest. A non-interactive explanation would simply be the output of a report. But both would need to be human-interpretable to be useful. Please clarify what you mean by "interactive" and "human interpretable".

Authors' response: We thank the reviewer for this helpful clarification. Our intended meaning of **interactive** agrees with the reviewer's interpretation: users can guide which explanations are displayed, for example by expanding or collapsing reasoning steps or requesting clarification. We have revised the description of Y2 in Section 3.3.2 accordingly.

We also clarify that **human-interpretable** describes explanations whose reasoning steps and connection to the result can be understood by the intended users; it is not introduced as an additional taxonomy concept.

Text in the marked manuscript

Location in the marked manuscript: Section 3.3.2, page 28, lines 709–722.

Reviewer 1 – Comment R1.32

3.3. Z: Evaluating: What is an integration level? What does it mean in practice for one method to be more integrated than another? Can you give examples?

Authors' response: We thank the reviewer for highlighting this ambiguity. We agree that **integration level** was inappropriate because the metric does not measure how closely neural and symbolic components interact within a method. Such an interpretation would require explicit architectural criteria that our concept annotations do not establish.

We have therefore replaced **integration level** with **Z3: Joint coverage**. As explained in the first paragraph of Section 3.4, this concept measures the combined breadth of capability and assurance concepts assigned to a paper. A higher Z3 value therefore indicates greater joint coverage under the revised metric, rather than tighter neural–symbolic integration. Section 6.1 provides the formal definition and specifies how Z3 is computed in Equation (25).

Text in the marked manuscript

Location in the marked manuscript: Section 3.4, pages 29–30, lines 746–772, lines 740–745; Section 6.1, Equation (25).

Reviewer 1 – Comment R1.33

4. Existing Surveys

This section is very interesting and eye-opening but it felt quite long considering the abstract puts focus on the contents of sections 5 and 6. Admittedly I'm not sure what to

suggest. Can it be compressed or moved to a later point in the survey?

Authors' response: We thank the reviewer for this thoughtful comment. We acknowledge that Section 4 is lengthy, and we have improved its readability to make the discussion easier to follow. However, we consider its current position important because it establishes the context for Sections 5 and 6 and explains how our methodology relates to existing NeSy surveys.

We have retained the detailed explanations to provide, in one place, a consolidated account of the survey methodologies identified through our literature search. Further compression could obscure differences in their scope, classification criteria, and evaluation approaches, while moving the section later would separate this methodological context from the analysis it supports. This consolidated account is intended to support both human readers and AI-assisted analysis of the NeSy survey literature.

Text in the marked manuscript

Location in the marked manuscript: Section 4.

Reviewer 1 – Comment R1.34

Tables 2/3: It might be interesting to include the dates of the surveys and to order by date to support later comments about trends with respect to alignment with the dimensions of the taxonomy.

Authors' response: We thank the reviewer for this helpful suggestion. The publication years are already included in the citations provided in the third column, "Survey paper," of Tables 14 and 15. We have retained this format because adding a separate year column would cause the tables to exceed the page margins. Readers can therefore identify the publication years directly from the existing citations. Furthermore, the rows are grouped by taxonomy name, and publications from different years may belong to the same group. A strictly chronological ordering would disrupt this grouping and make comparisons within each taxonomy less clear. We have therefore retained the existing organization.

Text in the marked manuscript

Location in the marked manuscript: Section 4, page 31, Table 14; Section 4, page 32, Table 15.

Reviewer 1 – Comment R1.35

For table 3 ID 18: Ref [123] was not the original proponent of the ADT taxonomy (<https://www.sciencedirect.com/science/article/abs/pii/0950705196819204>), which dates back as far back as 1995 (during the pioneering era of section 5.2). I'd recommend including the original in the references and some comments on the same in the main text. [123] borrows the taxonomy to frame new ideas from an earlier perspective. Please address this where [123] is discussed in the text also.

Authors' response: We thank the reviewer for identifying this important attribution issue. We have revised Table 15, ID 18, and the corresponding discussion in the main text to make clear that [Townsend et al. \(2019\)](#) was not the original proponent of the ADT taxonomy. The original source, Andrews, Diederich, and Tickle [Andrews et al. \(1995\)](#), has now been added to the references and to Table 15, ID 18.

Changes made: The revised manuscript now explains that [Townsend et al. \(2019\)](#) adopts the earlier ADT taxonomy of rule-extraction and rule-refinement techniques, whereas the taxonomy itself was originally proposed by Andrews, Diederich, and Tickle in 1995 [Andrews et al. \(1995\)](#). We also clarified in the main text that the original ADT taxonomy belongs to the pioneering period discussed in Section 5.3. This correction avoids attributing the taxonomy to the later survey and clarifies that [Townsend et al. \(2019\)](#) uses the earlier taxonomy retrospectively to frame newer work on extracting relational explanations from deep neural networks.

Text in the marked manuscript

Location in the marked manuscript: Section 4, Page 45, lines 1239–1249.

Reviewer 1 – Comment R1.36

5. Literature Survey

Please elaborate on how the 43 papers selected, and why only 43? There’s a lot of NeSy methods out there

Authors’ response: We thank the reviewer for this important comment. The original set of 43 papers was not intended as a predetermined or exhaustive limit. Rather, it constituted the corpus identified and evaluated using the initial search procedure. To reassess and broaden the coverage, we developed a structured search prompt organised according to the newly defined publication periods, grouped by decade from 1940–1949 through 2020–2026, and repeated the search using the GPT-6-Astra OpenAI web-search tool. The complete prompt, including the inclusion and exclusion criteria used to guide the publication search, has been added to Appendix A.2, entitled **Prompt for the publication search**. We subsequently screened the retrieved records, checked their bibliographic metadata and peer-reviewed publication status, and checked for duplicates and different versions of the same contribution. A number of candidates were excluded. We also reassessed papers considered in previous submission of this survey and excluded the following six: [Sun et al. \(2020\)](#); [Hölldobler and Kalinke \(1999\)](#); [Cappart et al. \(2023\)](#); [Cameron et al. \(2020\)](#); [Hatzilygeroudis and Prentzas \(2000\)](#); [Tsakonas et al. \(2004\)](#). The [reference list](#) provides the reasons for excluding each of these six references. The final publication list contains 181 distinct citation keys used in the **Literature Survey** section. We also revised the **Acknowledgements** section to disclose that GPT-6-Astra OpenAI, including its web-search functionality, was used to support the publication search.

Text in the marked manuscript

Location in the marked manuscript: Appendix A.2; Appendix A1; Section 5, lines 1265–1288.

Reviewer 1 – Comment R1.37

Please remind the reader why Z is excluded from the tables if section 4 remains where it is at its current length.

Authors’ response: We thank the reviewer for pointing out that the role of the Z dimension should be clarified before the survey tables. We have added an explicit clarification at the end of Section 3.4 explaining that Z is not used as a direct annotation dimension in Sections 4 and 5. Instead, Z is computed later in Section 6 from the completed X and Y annotations and auxiliary integration information.

Changes made: A clarifying sentence was added at the end of Section 3.4 to remind readers

that Sections 4 and 5 use only the X and Y dimensions for describing and annotating papers and frameworks, while the Z dimension is reserved for the subsequent evaluation step in Section 6.

Text in the makred manuscript

Location in the marked manuscript: Section 3.4, pages 29–30, lines 754–762.

Reviewer 1 – Comment R1.38

As mentioned above please include more commentary on transitions between eras (extends to defending choice of name).

Authors’ response: At the beginning of each decade-based survey subsection, we now discuss transitions between the decades, and papers reviewed in the preceding period and those reviewed in the current subsection. These comparisons highlight continuity and changes in reported capabilities, assurance and ANN architectures. We also use explicit publication periods as subsection titles to avoid implying that a particular architecture was dominant or that a characteristic, such as differentiability, was absent in earlier periods.

Text in the makred manuscript

Location in the marked manuscript: Section 5.2, lines 1304–1311; Section 5.3, lines 1330–1343; Section 5.4, lines 1501–1511; Section 5.5, lines 1615–1623; Section 5.6, lines 1886–1900.

Reviewer 1 – Comment R1.39

For example, is the message that during 2023-2024 graph-based models were the dominant paradigm for NeSy? Were models before 2018 not differentiable?

Authors’ response: **We thank the reviewer for highlighting this ambiguity.** In the revised survey, which covers 191 publications, Tables 30 and 31 in the “2020–2026” subsection report the X – Y annotations and ANN architectures of NeSy frameworks published in 2023 and 2024, respectively. These tables show that HNN-based models, rather than GNN-based models, are the most frequently represented architectures in the surveyed publications for these years. This observation concerns our corpus of PL-focused frameworks and should not be interpreted as a claim about the dominant paradigm across NeSy research as a whole. GNNs were assigned as an ANN architecture label to 12 publications published between in 2023 and 2024.

We also clarify that models published before 2018 were not necessarily non-differentiable. HNN-based models were the most frequently represented architectures in the surveyed publications from this period (see Section 5).

Text in the makred manuscript

Location in the marked manuscript: Section 5, Tables 30-31.

Reviewer 1 – Comment R1.40

5.1. 1943-1989: Foundational neural logic era: While I absolutely agree it’s important to address McCulloch and Pitts but I’m not sure we can call this an ‘era’ when it’s just one paper significantly earlier than the pioneering era begins.

Authors’ response: We agree that a single early publication does not justify describing the period as a distinct “era.” The work of McCulloch and Pitts (1943) is now described and annotated in Section 5.1, “1940–1979,” without the previous era designation.

The introduction to Section 6 explains why this reference is excluded from the evaluation. Section 6.2, “Data Preprocessing,” explicitly states that the paper is retained for historical context but excluded from the canonical dataset and quantitative evaluation. Its inclusion in the historical discussion therefore does not imply its inclusion in the quantitative analysis.

Text in the makred manuscript

Location in the marked manuscript: Section 5.1; Section 6, lines 2742–2764; Section 6.2, lines 2825–2838.

Reviewer 1 – Comment R1.41

5.4. 2018-2020: Differentiable Neuro-Symbolic AI Era: Where does the feature of differentiability come into this? "differentiable" doesn't come up until the last paragraph of the page. Furthermore, older models were differentiable. KBANN and CILP (Pioneering era) were both differentiable. Why do differentiable models stand out during 2028-2020 in particular?

Authors’ response:

We agree that differentiability is not exclusive to models published during 2018–2020 and therefore does not justify the previous period designation. We have revised the temporal grouping criteria and now organize publications into ten-year periods, rather than eras defined by architectural or methodological characteristics. The term “differentiable” is no longer used to group publications. The revised grouping is presented in Table 21, referenced in the introduction to Section 5.

Text in the makred manuscript

Location in the marked manuscript: Section 5, Table 21.

Reviewer 1 – Comment R1.42

5.6. 2023–2024: Graph-based: A survey on Graph Neural Networks for NeSy was published back in 2020 (differentiable era): <https://arxiv.org/pdf/2003.00330>. What makes graph-based models stand out during 2023-2024?

Authors’ response: We thank the reviewer for this important observation. We agree that graph-based NeSy models predate 2023–2024, and their use does not justify defining these years as a distinct graph-based era. In the revised survey, publications are grouped into ten-year periods rather than architecture-based eras.

Regarding the question, “What makes graph-based models stand out during 2023–2024?”, the surveyed frameworks illustrate their ability to exploit relationships between variables and clauses, support diverse reasoning tasks, and, in some cases, generalize to larger problems. In 2023, their applications include MaxSAT optimization, satisfiability prediction, and UNSAT-core prediction. The surveyed publications from 2024 also include CDCL solver guidance, adaptive solver selection, and cryptographic SAT applications. NeuRes is particularly notable for generating checkable resolution proofs of unsatisfiability. Most other graph-based frameworks surveyed for these years provide approximate predictions or heuristic assistance, including search-space

pruning, without reported proof-certification mechanisms. These observations describe the contributions represented in our corpus, rather than establishing that graph-based models originated or became dominant during this period.

Text in the marked manuscript

Location in the marked manuscript: Section 5.6, Tables 30-31.

Reviewer 1 – Comment R1.43

Evaluation of Existing Work

I feel it is difficult to include the foundational era in the plots here. With one paper so early on with respect to the busy activity in the other eras, it's really more of an outlier. Don't get me wrong though, it's a very important paper and It's nice to see it discussed earlier.

Authors' response: Please see our response to Comment [R1.40](#).

Reviewer 1 – Comment R1.44

Note my earlier suggestion to plot against publication date. I appreciate there would probably still be a downward trend, and my intuition from reading so many NeSy papers down the years would lead me to agree with the observation anyway. I think nobody would be able to argue with the trend if it is presented at the finest possible granularity, and the authors have the data to plot it I think.

Authors' response: In Section 6.5, we now present publication-year plots of the arithmetic means, medians, and binned Shannon entropy of Z_1 , Z_2 , and Z_3 across the evaluated publications (Figures 9, 10, and 11, respectively). The bottom panel of each figure shows the number of evaluated publications per year, allowing readers to assess the annual trends alongside the corresponding publication counts. These figures were generated using an automated pipeline implemented as a Python project, accessible through Listing 1 in Appendix A1.

Text in the marked manuscript

Location in the marked manuscript: Section 6.5, Figures 9–11.

Reviewer 1 – Comment R1.45

When discussing the median, what is meant by "central" level of reasoning?

Authors' response: We thank the reviewer for highlighting this ambiguity. In the revised manuscript, we no longer associate the mean or median with a “central level of reasoning.” Instead, these statistics summarize the numerical Z scores across the evaluated publications; they do not denote an ordinal reasoning level.

Reviewer 1 – Comment R1.46

Regarding the decline 2025 for $Z_{1.1}$, I wonder if the possible effect of citation lag has been considered? Literature that changes the result for 2025 may exist, but because it is new, authors may not be aware of it yet (e.g. because it's too new for other authors to have also discovered and cited it yet). Authors may want to highlight this caveat before drawing conclusions about the numbers for 2025 as a whole.

Authors’ response: In the revised survey, we searched for publications from 2025 and 2026 using a structured prompt with GPT-6-Astra OpenAI’s web search tool, with the systematic search completed on 08 October 2026. The structured prompt used for this search is provided in Appendix A2. Table 32 in Section 5.6, “2020–2026,” lists publications from 2025 included in our survey.

Although this updated search extends the coverage, it cannot eliminate possible omissions caused by citation lag, indexing delays, or limited discoverability of recent publications. Consequently, the results for 2025 should be interpreted as observations about the surveyed corpus, rather than conclusive evidence of a decline across all relevant research published that year.

Text in the marked manuscript

Location in the marked manuscript: Section 5.6, Table 32; Appendix A2

Reviewer 1 – Comment R1.47

Conclusions

This section explains what the survey did but unfortunately lacks any commentary on what it concludes. What are the take-home messages? What does the survey show is missing from other surveys and the technologies themselves? How should NeSy researchers take these findings and change how they critique their own and other researcher’s work?

Authors’ response: We have revised the Conclusion to state the main findings and take-home messages explicitly, including gaps in the coverage of PL reasoning, proving, and assurance in existing surveys and frameworks. We also discuss how NeSy researchers can use these findings to assess their own and others’ work, by distinguishing demonstrated capabilities from claimed guarantees and examining the evidence supporting correctness, proof validation, and certification.

Text in the marked manuscript

Location in the marked manuscript: Section 7.

Reviewer 1 – Comment R1.48

I think the authors pick on something quite important in this paper but unfortunately don’t give it the emphasis it deserves in the conclusions: that reasoning and proving capabilities and assurance in propositional logic (PL) remains limited in modern NeSy models and surveys, perhaps because of such an emphasis on explainability in the past decade, and more attention needs to be given to these capabilities.

Authors’ response: We thank the reviewer for emphasizing this important point. We have strengthened the discussion in the Introduction, Section 6.5, and the Conclusion sections to highlight the limitations in PL reasoning and proving capabilities and assurance documented in the surveyed literature, and the need for greater research attention to these aspects. We also clarify that explainability alone does not establish correctness or replace proof validation and certification. Although the emphasis on explainability may have contributed to this imbalance, our findings do not establish a causal relationship.

Text in the marked manuscript

Location in the marked manuscript: Section 1, Section 6.5, Section 7.

Reviewer 1 – Comment R1.49

Also is there scope to extend to predicate logics, temporal logics, and other type of logic in general?

Authors' response: The present survey is deliberately restricted to propositional logic (PL), because our goal is to analyze a focused and under-discussed part of NeSy capability and assurance mechanisms, especially SAT/UNSAT solving, proof search, witness generation, refutation, and certification. Extending the survey to predicate logic, temporal logic, description logic, modal logic, or other logics would substantially change the scope, since each logic has different syntax, semantics, reasoning tasks, proof systems, decidability properties, and assurance mechanisms. Covering these additional logics would therefore require a separate literature analysis and likely additional taxonomy dimensions. We now clarify that such an extension is valuable but beyond the scope of the present PL-focused survey and is left for future work.

Please see our response to Comment [R1.21](#), which discusses the scope for extending the taxonomy to predicate, temporal, and other logics.

Response to Reviewer 2

We thank a Reviewer 2 for carefully examining the configuration and presentation of the taxonomy and for identifying several points that required further explanation and restructuring. To make the revision process transparent, we provide a marked manuscript in which text removed in response to the reviewer's comments is shown in red with strikethrough, while newly added text is shown in **Green** and **Violet**. The revised manuscript contains only the final updated text.

Reviewer 2 – Comment R2.1

This paper presents a detailed survey and taxonomy (called X-Y-Z taxonomy) of NeSy frameworks for reasoning and proving in PL. The basic (X-Y) taxonomy concerns two levels/criteria: capability and assurance. Taxonomy is formalized as an ontology based on the SKOS standard. The authors also introduce an automated methodology for evaluating NeSy frameworks across reasoning and proving capability and assurance, which also is implemented in Python. Based on that, statistical results for various time periods are provided.

Authors' response: We thank the reviewer for this clear summary of our work and for recognizing its main contributions: the X-Y-Z taxonomy, its SKOS-based implementation, and the automated evaluation methodology. We appreciate the reviewer's careful reading and constructive feedback, which have helped us improve the clarity and precision of the manuscript.

Reviewer 2 – Comment R2.2

The authors have done a huge job in reviewing NeSy papers and also a very detailed work on configuring the taxonomy. However, I think that there are several configuration and presentation problems. My main concerns are as follows:

* The authors talk about "Reasoning and Proving in PL". Why in PL? Which other options are there and why they have been ignored?

Authors' response: We thank the reviewer for recognizing the scope of our work and for raising this important question. We focus on PL to maintain a clearly defined scope for comparing NeSy reasoning and proving capabilities and assurance mechanisms. The revised survey includes 181 publications, illustrating the breadth of the literature considered within this scope.

Other options include predicate, temporal, and modal logics. These have not been dismissed; their additional constructs and reasoning requirements would require extending the taxonomy and the survey methodology. Please see our responses to Comments [R1.21](#) and [R1.49](#) for further discussion of this scope decision and possible extensions.

Reviewer 2 – Comment R2.3

* The abstract should be more informative.

Authors' response: We have revised the Abstract to describe the survey's scope, methodology, main findings, and contributions more explicitly, including the X-Y-Z taxonomy and the automated evaluation of NeSy frameworks for PL tasks. A table of abbreviations used throughout the paper has been added immediately after the Abstract.

Text in the marked manuscript

Location in the marked manuscript: Abstract

Reviewer 2 – Comment R2.4

*Introduction is too short and not adequately informative. The authors talk directly from the beginning about "a critical gap", which however does not come out from some kind of "exploration" from existing works. There are no references at all in the introduction. Also, no specific motivations and objectives are stated.

Authors' response: We thank the reviewer for this important comment, which we used as a guideline for revising the Introduction. The revised section includes explicit motivation and objectives, relevant references, detailed contributions, an overview of the survey, and an explanation of the novelty and scope of this synthesis. It also places the identified research gap in the context of existing work and explains why a focused examination of PL reasoning, proving, and assurance is needed.

Text in the marked manuscript

Location in the marked manuscript: Section 1, Section 7.

Reviewer 2 – Comment R2.5

* In the Background section, a long part is devoted to PL and only a very short section (Section 2.6) to Neural Networks. It should be adequately expanded. Also, I think Section 2.6 should be moved before Section 2.5.

Authors' response: We thank the reviewer for this helpful suggestion. We have expanded the neural network background, now presented in Section 2.5, as described in our response to Comment [R1.25](#). We have also exchanged the order of the former Sections 2.5 and 2.6, placing the neural network introduction first, as discussed in our response to Comment [R1.22](#).

Text in the marked manuscript

Location in the marked manuscript: Section 2.5, Section 2.6.

Reviewer 2 – Comment R2.6

* The X dimension taxonomy (Section 3.1) is very detailed. The question is whether so many details are useful; what do they offer? As it is clear from Tables 2, 3 and 18, only about one third of the configured categories correspond to real NeSy frameworks. What about the rest? Do they show real possibilities for new types of NeSy frameworks?

Authors' response: We thank the reviewer for these important questions. We have revised the X–Y–Z taxonomy to improve its clarity and organization. The detailed X concepts distinguish specific reasoning and proving capabilities that broader categories could obscure. Their purpose is to support precise annotation and comparison, rather than to suggest that each concept corresponds to a distinct type of NeSy framework.

Section 6.3, “Concept Frequencies and Weights,” now reports the numbers of used and unused concepts and explains why some concepts are not assigned. Table 33 presents corpus-wide frequencies and weights of X and Y concepts across the 178 evaluated publications. Frequency denotes the number of publications assigned a concept, and its weight is that frequency divided by 180, rounded to three decimal places.

Text in the marked manuscript

Location in the marked manuscript: Section 6.3, pages 93–95.

An unassigned concept does not necessarily indicate an unrealized capability: it may serve an organizational role in the hierarchy, or the surveyed publications may not provide sufficient evidence for its assignment. Such concepts can identify potential directions for further investigation, but their absence from the annotations does not establish either the feasibility of a new framework or the absence of relevant work outside our corpus. We therefore distinguish the taxonomy’s conceptual coverage from the capabilities documented in the evaluated publications.

Text in the marked manuscript

Location in the marked manuscript: Section 6.3.

Reviewer 2 – Comment R2.7

* I think that combined terms for description of leaf categories is not a good policy, as for example, X1.3.1.2.1: Meta-reasoning and strategy adaptation. If the two conjuncts (meta-reasoning, strategy adaptation) are of similar nature, give a unified term name, otherwise split it in two sub-categories. Also, long descriptive terms, like e.g. X1.3.1.1.1: Core propositional SAT solving, should be shortened (e.g. make it SAT solving).

Authors’ response: We thank the reviewer for this helpful recommendation. In response, we systematically revised the X taxonomy by splitting combined terms when they represent distinct capabilities, replacing closely related conjuncts with a single unifying term, and shortening long or parenthetical labels. For example, **Meta-reasoning and strategy adaptation** was split into **Meta-reasoning** and **Strategy adaptation**, while **Core propositional SAT solving** was shortened to **SAT solving**. We also adopted singular noun forms consistently because each concept represents an instance of SKOS ontology class. These revisions make the concept labels more concise, precise, and internally consistent.

The implementation of the revised taxonomy is available in the `taxonomy` folder of the Python project, accessible through Listing 1 of Appendix A1.

Changes made:

Exact changes made to the taxonomy

Here is a complete, itemized list of every modification, with the original wording shown for reference. First we revised

1. X0

Original: X1: Reasoning and proving capability level

Changed to: X0: Capability level

Reason: The label was shortened by removing the term "Reasoning and proving". The revised term still covers the intended scope.

Text in the marked manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

2. X3

Original: X1.3: Strategic reasoning and adaptation

Changed to: **X3: Solution strategy**

Reason: The original label combined two related notions. The narrower concepts in this branch, including logical strategy, reasoning method, SAT solving, resolution, heuristic adaptation, and MaxSAT optimization, are more coherently unified under the shorter parent concept **solution strategy**.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

3. **X4**

Original: **X1.4: Search and control**

Changed to: **X4: Search strategy**

Reason: The original label used a conjunctive construction. The narrower concepts in this branch, including lemma discovery, clause selection, search-space pruning, branching heuristic, SAT heuristic optimization, and lemma reuse, all describe mechanisms for guiding or improving search. They are therefore unified under the shorter concept **Search strategy**.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

4. **X5**

Original: **X1.5: Training and robustness**

Changed to: **X5: Learning robustness**

Reason: The original label combined training with robustness. The revised label removes the conjunction and better captures the common scope of this branch, namely learning-related mechanisms that support robust reasoning behaviour, such as logical-reasoning generalization, paradox-resistant training, and UNSAT prediction.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

5. **X6**

Original: **X1.6: Integration and expressivity**

Changed to: **X6: Architectural capability**

Reason: The original label combined two different design-level aspects. The narrower concepts, including algorithmic alignment and logic expressivity, both describe architectural properties of a NeSy framework. The revised label therefore provides a shorter and more general parent concept.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

6. **X8**

Original: **X1.8: Robustness and Interpretability**

Changed to: **X8: Reliability**

Reason: The original label combined two related properties. The revised label unifies

robustness and interpretability under the broader capability-side concept **Reliability**. This keeps the branch in the *X* taxonomy, because it describes properties of the reasoning capability itself, whereas the *Y* taxonomy captures assurance mechanisms used to verify, certify, or explain that capability.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

7. **X1.1**

Original: X1.1.1.1.1: Decision-only (SAT/UNSAT)

Changed to: X1.1: Satisfiability decision

Reason: Shortened the long descriptive label and removed the parenthetical phrasing.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2.1, page 18, lines 519–526.
Page 19, see new X1.1: Satisfiability decision concept added to Figure 3.

8. **X1.2**

Original: X1.1.2: Logical inference (derivation of relations)

Changed to: X1.2: Logical consequence

Reason: Removed the lengthy parenthetical explanation to make the term concise.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, see new X1.2: Logical consequence concept added to Figure 3.

9. **X2 → X2.1 / X2.2**

Original:

X1.2: Witness construction

X1.2.1: Constructive witness generation

Changed to:

X2: Constructive evidence generation

X2.1: Satisfiability witness generation

X2.2: UNSAT refutation

New broader concept is X2: Constructive evidence generation. The revised branch distinguishes two complementary forms of constructive evidence:

- X2.1: Satisfiability witness generation concerns generating a witness, such as a truth assignment, showing that a PL formula is satisfiable.
- X2.2: UNSAT refutation concerns constructing a proof or certificate showing that a PL formula is unsatisfiable.

Reason: The original concept Constructive witness generation was too narrow because it mainly captured evidence for satisfiability. The revised parent concept X2: Constructive evidence generation generalizes this branch to cover constructive evidence for both satisfiable and unsatisfiable cases. The original witness-generation meaning is retained as the more specific concept X2.1: Satisfiability witness generation,

while **X2.2: UNSAT refutation** is moved here because it also produces constructive evidence, but in the form of proof evidence for unsatisfiability.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2.2, page 18, lines: 530–539; Page 19, Figure 3.

10. **X3.1**

Original: X1.3.1: Logical strategies

Changed to: X3.1: Logical strategy

Reason: Adopted singular noun forms.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3; Section 3.2.3, pages 20–22.

11. **X3.1.1.1**

Original: X1.3.1.1.1: Core propositional SAT solving

Changed to: X3.1.1.1: SAT solving

Reason: The lengthy qualifier **Core propositional** was removed, yielding the concise leaf label SAT solving, exactly as requested by the reviewer for long descriptive terms.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, see new X3.1.1.1: SAT solving concept added to Figure 3; Section 3.2.3, page 20, lines 559–561.

12. **X3.1.2.1 / X3.1.2.4**

Original: X1.3.1.2.1: Meta-reasoning and strategy adaptation

Changed to:

- X3.1.2.1: Strategy monitoring
- X3.1.2.4: Strategy adaptation

Reason: No structural merge was performed (they remain distinct because they are not of identical nature). The reviewer’s example of a combined term is avoided. The label was shortened and the repeated term **reasoning** was removed. The revised label preserves the intended meaning because the concept denotes the capability of a NeSy framework to monitor and improve its own PL solution strategy during the solving process.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

13. **X3.1.2.2**

Original: X1.3.1.2.2: Reinforcement learning for clause selection

Changed to: X3.1.2.2: RL-guided clause selection

Reason: Shortened the descriptive label while preserving its essential meaning: the use of reinforcement learning to guide clause selection.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

14. **X4.1 / X4.6** Original: X1.4.1: Lemma discovery and reuse Changed to:

- X4.1: Lemma discovery
- X4.6: Lemma reuse

Reason: The original leaf combined two related but distinct capabilities. **Lemma discovery** concerns deriving useful intermediate results during reasoning, whereas **Lemma reuse** concerns selecting and applying previously established lemmas to guide or shorten subsequent search. Since either capability may be supported independently, the original combined concept was split into two sibling concepts, following the reviewer’s recommendation.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2.4, page 22, lines: 588–589, 606–607; Page 19, Figure 3.

15. **X1.4.3**

Original: X1.4.3: Pruning and control of search space

Changed to: X4.3: Search-space pruning

Reason: Unified the two similar conjuncts into one concise term and shortened the descriptive phrasing.

Location in the marked manuscript: Section 3.2, Figure 3.

16. **X1.5.1**

Original: General logical reasoning

Changed to: logical generalization

Reason: Unified the two similar conjuncts into one concise term and shortened the descriptive phrasing.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, Figure 3.

17. **X5.3 / X2.2**

Original: X1.5.3: Core UNSAT prediction and refutation

Changed to:

- X5.3: UNSAT-core prediction
- X2.2: UNSAT refutation

New broader concept for X2.2: UNSAT refutation is: X2: Constructive evidence generation

The two concepts represent related but distinct capabilities:

- X5.3: UNSAT-core prediction concerns predicting clauses or variables that are likely to belong to an unsatisfiable core, typically to guide a SAT solver.

- **X2.2: UNSAT refutation** concerns constructing a proof or certificate that establishes unsatisfiability.

Reason: The original label was a conjunctive leaf combining two distinct capabilities. Following the reviewer’s recommendation, the combined term was split. **UNSAT-core prediction** remains under **X1.5: Learning robustness**, because it denotes a learned prediction mechanism used to guide SAT solving. By contrast, **UNSAT refutation** was moved to **X1.2: Constructive evidence generation**, because it denotes the construction or validation of proof evidence for unsatisfiability rather than a learning-based robustness capability. The label **UNSAT-core prediction** was retained instead of the broader **UNSAT prediction**, because it more precisely reflects the cited work, where predictions concern unsatisfiable cores rather than merely classifying a formula as unsatisfiable.

Text in the marked manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

18. **X6.2**

Original: **X1.6.1.1.2 Logic expressivity & proof calculus support**

Changed to: **X6.2: Logic expressivity**

Reason: Removed the combined second conjunct (“& proof calculus support”) and shortened the label; the second idea was treated as secondary and dropped for conciseness (or could be split later if needed).

Text in the marked manuscript

Location in the marked manuscript: Section 3.2, Figure 3.

19. **X1.7 / X1.7.1 / X1.7.1.1**

Original:

X1.7: Verification support

X1.7.1: Certificate evaluation

X1.7.1.1: Soundness & verified guarantees

Changed to: **X7: Soundness guarantee**

Reason: The original branch mixed capability-side soundness with assurance-side verification and certificate checking. To remove this overlap, the *X* taxonomy now retains only **Soundness guarantee** as a capability-level property of the reasoning/proving framework. Certificate evaluation and proof checking are not kept as *X* sub-concepts, because they describe assurance mechanisms rather than reasoning capability itself; these mechanisms are represented in the *Y* taxonomy. Therefore, **X7: Soundness guarantee** does not require narrower *X* sub-concepts.

Text in the marked manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

20. **X8.2**

Original: **X1.8.2.1: Explainability & human interpretability**

Changed to: (leaf removed; parent becomes the leaf)

Structural change: The combined leaf was eliminated. The parent **X8.2: Interpretability**

now serves as the terminal concept.

Reason: The two conjuncts are essentially the same idea; a single unified term is preferred.

21. Minor consistency clean-ups applied to X taxonomy

- Removed all remaining ampersands (&) and long parenthetical expansions.
- Ensured every leaf is a single, short noun phrase rather than a multi-conjunct descriptive sentence.
- Kept hierarchy depth and numbering almost identical so the structure remains recognizable.

No other concepts were renamed, split, or merged. The rest of the taxonomy is identical to the original.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2

Reviewer 2 – Comment R2.8

On the other hand, what is the rationale of having categories having only one subcategory?

Authors' response: We thank the reviewer for this observation. Categories that contain only a single subcategory add hierarchical depth without providing discriminative or organizational value. In accordance with standard taxonomy design principles, we have systematically identified some (not every) unary branch and collapsed it by promoting the content of the unique child (or the terminal leaf) to the parent level. This produces a flatter, more economical structure while fully preserving the original conceptual distinctions.

Changes made: The following unary branches were eliminated by merging:

- X1.1

Original:

X1.1.1: Decision mechanisms

X1.1.1.1: Basic decisions

X1.1.1.1.1: Decision-only (SAT/UNSAT)

Collapsed into a node X1.1: Satisfiability decision.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

- X2

Original:

X1.2: Witness construction

X1.2.1: Constructive witness generation

Collapsed into a node X2: Constructive witness generation.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

- X3.2

Original:

X1.3.2: Optimization extensions

X1.3.2.1: MaxSAT optimization

Collapsed into X3.2: MaxSAT optimization.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

- X6.1 / X6.2

X1.6.1: Logical integration

X1.6.1.1: Logic mechanisms

X1.6.1.1.1: Algorithmic alignment

X1.6.1.1.2: Logic expressivity & proof calculus support

The two children of Logic mechanisms are now direct children of X6:

X6: Architectural capability

X6.1: Algorithmic alignment

X6.2: Logic expressivity

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

- X7

X1.7: Verification support

X1.7.1: Certificate evaluation

X1.7.1.1: Soundness & verified guarantees

Collapsed into X7: Soundness guarantee.

Text in the makred manuscript

Location in the marked manuscript: Section 3.2, page 19, Figure 3.

No other structural changes were required at this point.

Text in the marked manuscript

Location in the marked manuscript: The revised taxonomy, in which these concepts have been consolidated, is presented in Figure 3, Section 3.2, page 19.

Reviewer 2 – Comment R2.9

* In Section 3.1, while the authors describe taxonomy categories in plain unstructured text, for one specific category (X1.3) and its subcategories they use numbered bullets. I cannot find any reason for not having a unified description way. This also holds for the descriptions in Section 4.

Authors' response: We have revised Sections 3.1 and 4 to use a uniform prose format without bullet points or numbered lists, including the description of X3 concept and its narrower concepts.

Text in the marked manuscript

Location in the marked manuscript: Section 3, Section 4.

Reviewer 2 – Comment R2.10

* Similar things to above should be also checked for Y dimension.

Authors' response: We thank the reviewer for this observation. We checked the Y taxonomy using the same revision principles applied to the X taxonomy. In particular, we shortened long labels, removed conjunctive or ambiguous terms, split combined concepts where the components represented distinct assurance functions, and collapsed one-child branches where the intermediate concept did not add an independent semantic level. As a result, the revised Y taxonomy has a more compact structure, clearer labels, and fewer unnecessary hierarchy levels, while preserving the distinction between reasoning/proving capability in the X dimension and assurance mechanisms in the Y dimension.

Changes made:

1. Y0

Original: Y1: Reasoning and proving assurance level

Changed to: Y0: Assurance level

Reason: The label was shortened by removing the term "reasoning and proving." The revised term still covers the intended scope.

Location in the marked manuscript: Section 3.3, page 26, Figure 4.

2. Y1

Original: Y1.1: Verification and certification

Changed to: Y1: Verification

Reason: The label was shortened by removing the term "reasoning and proving." The revised term still covers the intended scope.

Location in the marked manuscript: Section 3.3, page 26, Figure 4.

3. Y1.1 / Y1.1.1 / Y1.1.2

Original:

- Y1.1.1: Certificate support
 - Y1.1.1.1: Certificate format support (DRAT/DRUP)
 - Y1.1.1.2: Provable soundness via external verifier

Changed to:

- Y1.1: Certificate support
 - Y1.1.1: Certificate format
 - Y1.1.2: External verification

Reason: The parent concept Y1.1.1: Certificate support was retained, but the two narrower labels were shortened and clarified. Y1.1.1.1: Certificate format support (DRAT/DRUP) was revised to Y1.1.1: Certificate support by removing the repeated word support and moving the DRAT/DRUP examples to the concept description. Y1.1.1.2: Provable soundness via external verifier was revised to Y1.1.2: External verification to avoid overlap with X7: Soundness guarantee and to emphasize the assurance mechanism: checking certificates or proofs by an external verifier.

Text in the makred manuscript

Location in the marked manuscript: Section 3.3, page 26, Figure 4.

4. Y1.2 / Y1.3

Original:

- Y1.1.2: Proof validation
 - Y1.1.2.1: Proof compression

Changed to:

- Y1.2: Proof validation
- Y1.3: Proof compression

Reason: The original hierarchy placed Y1.1.2.1: Proof compression as the only narrower concept under Y1.3: Proof compression. To address the reviewer's concern about one-child branches, the intermediate dependency was removed and Y1.3: Proof compression was promoted to a direct concept under Y1: Verification. This keeps both assurance functions visible while avoiding unnecessary hierarchy depth.

Text in the makred manuscript

Location in the marked manuscript: Section 3.3, page 26, Figure 4.

5. Y1.4 / Y1.4.1

Original:

- Y1.1.3: Refutation certification
 - Y1.1.3.1: Refutation or UNSAT certification

Changed to:

- Y1.4: Refutation certification
 - Y1.4.1: UNSAT certification

Reason: The concept was renumbered because Proof compression was promoted from Y1.1.2.1 to Y1.3. The leaf label Refutation or UNSAT certification was shortened to UNSAT certification to remove the disjunctive construction and avoid repeating the

broader parent term **Refutation certification**. The revised label more directly describes the assurance function: certifying that an unsatisfiability result is supported by a checkable refutation or certificate.

Text in the makred manuscript

Location in the marked manuscript: Section 3.3, page 26, Figure 4.

6. Y2

Y1.2: Explainability for assurance

Y1.2.1: Interactive explainability

Collapsed into a single concept Y2: **Interactive explanation**.

Reason: This revision also addresses comment [R2.8](#) concerning the rationale for categories with only one subcategory. The intermediate parent **Explainability for assurance** concept was removed because it introduced an unnecessary one-child branch, while the revised concept **Interactive explanation** concept directly captures the assurance mechanism represented in this part of the taxonomy.

Text in the makred manuscript

Location in the marked manuscript: Section 3.3.2, page 28, lines 711–724.

Changes made: The following unary branches were eliminated by merging:

7. Y1.3

Y1.3: Core soundness guarantee

Y1.3.1: Soundness

Collapsed into Y3: **Soundness assessment**.

Reason: The original branch introduced a one-child hierarchy and overlapped with the X-taxonomy concept X7: **Soundness guarantee**. To preserve the distinction between X and Y, the revised Y concept is named **Soundness assessment**: it captures the assurance-side activity of assessing or checking soundness, while X7: **Soundness guarantee** remains the capability side property.

Text in the makred manuscript

Location in the marked manuscript: page 26, Figure 4.

Reviewer 2 – Comment R2.11

* In Section 4 (second paragraph), the authors state that they traced 76 NeSy papers between 1971 and 2025, from which 25 are survey/review papers. So the rest are $76 - 25 = 51$ non-survey papers. In Section 5 (first paragraph) they state that "This section surveys 43 research papers published between 1943 and 2025. This means that some $(51 - 43 = 8)$ papers are missing or there is a miscalculation. This should be clarified.

Authors’ response: We thank the reviewer for identifying this ambiguity. The 76 papers refer to the results retrieved by a SPARQL query executed against a SPARQL endpoint. From these results, we selected 26 survey/review papers for discussion in Section 4, pages 31–32, Tables 14 and 15. The remaining 50 papers were not used in this work; they were not the source pool from which the 43 research papers reported in the original Section 5 were selected. Therefore, subtracting 43 from 50 does not identify missing papers, as these counts refer to different selection processes.

The SPARQL query and its results are available in the `sparql` directory of the Python project, accessible through Listing 1 in Appendix A1.

Reviewer 2 – Comment R2.12

* In Section 4, a set of 25 survey/review papers are described in a plain and unstructured way from the start to the end. On the other hand, descriptions are so detailed (in several cases information about specific pages and tables are given). These make the reader to be lost after a while and not focus on more abstract concepts, which I think should be the objective. The authors, should create at least one-level subsections of Section 4, probably based on the existing taxonomies presented in Tables 2 and 3, possibly combining some of them in one subsection. In this context, dedicate first subsection to Kautz’s and Yu et al’s taxonomies. By the way, remove "(continue)" from the title of Table 3 and insert "(continue in Table 3)" at the end of the title of Table 2.

Authors’ response: We thank the reviewer for these helpful suggestions. We have slightly revised Section 4 to improve its readability and make the discussion more concise, reducing excessive detail and helping readers focus on the main concepts and relationships between the reviewed surveys. These revisions also address the similar concerns about the length and focus of this section raised in comment [R1.33](#).

Text in the marked manuscript

Location in the marked manuscript: Section 4.

Reviewer 2 – Comment R2.13

* In Section 5, first paragraph: "as illustrated in Figure 9" → "as illustrated in Table 9" (Figure 9 is too far). Also, [56] is missing from Table 11.

Authors’ response: We thank the reviewer for identifying these issues. We have corrected “Figure 9” to “Table 21” in the first paragraph of Section 5.

We have also added [Hatzilygeroudis and Prentzas \(2000\)](#), previously reference [56], to the exclusion list, with an explanatory note accompanying its bibliography entry.

This reference remains cited in Table 20 of Section 4 when discussing an existing survey’s classification: within category K–4, the authors classify the work in [Hatzilygeroudis and Prentzas \(2000\)](#) as a NeSy framework without reasoning support. Its inclusion in that discussion does not imply its inclusion among the frameworks evaluated in our survey.

Text in the marked manuscript

Location in the marked manuscript: Section 4, page 41, Table 20

Reviewer 2 – Comment R2.14

* In Section 6, the authors mention that "A higher value of a indicates a higher reasoning capability level (for X) or a higher proving assurance level (for Y). For example $F1(X1.6.1.1.2) = 6$ and $F2(Y 1.1.2) = 1$ ". So, it means that $F1(X1.6.1.1) = 6$ hence that X1.6.1.1.2 and X1.6.1.1 are of the same reasoning capability level, although X1.6.1.1 is higher in the taxonomy-ontology?

Authors' response: We thank the reviewer for identifying this ambiguity. We have revised the methodology and evaluation of NeSy frameworks. The numerical Z scores are now computed from corpus-wide concept frequencies and framework coverage, rather than numerical components of taxonomy identifiers.

Concept identifiers do not represent ordinal capability or assurance levels. A parent concept is broader than its child, but its position does not imply greater reasoning capability or stronger assurance. Consequently, we no longer use identifier indices to compare or rank concepts. Please also see our response to the related Comment [R1.8](#).

Concluding Statement

We again thank the editor and reviewers for their constructive and encouraging comments. Their recommendations have helped us improve the conceptual foundations, methodological transparency, organization, and presentation of the manuscript. We hope that the revised manuscript now addresses all the concerns raised and is suitable for publication in [Neurosymbolic Artificial Intelligence](#) journal.

Sincerely,

Nenad Krdzavac and Felix Engel

References

- Robert Andrews, Joachim Diederich, and Alan B. Tickle. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 8(6): 373–389, 1995. ISSN 0950–7051. doi: [https://doi.org/10.1016/0950-7051\(96\)81920-4](https://doi.org/10.1016/0950-7051(96)81920-4). URL <https://www.sciencedirect.com/science/article/pii/0950705196819204>.
- Chris Cameron, Rex Chen, Jason Hartford, and Kevin Leyton-Brown. Predicting propositional satisfiability via end-to-end learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3324–3331, 2020. doi: 10.1609/aaai.v34i04.5733. URL <https://ojs.aaai.org/index.php/AAAI/article/view/5733>. **Excluded:** The exchangeable and message passing architectures predict satisfiability status from CNF representations. Although extensions jointly predict variable assignments, the paper does not demonstrate that these satisfy every clause. Sections 3–5 report prediction methods and accuracy, without demonstrating a PL reasoning or proving mechanism or direct guidance of a symbolic SAT solver. The paper therefore falls under the exclusion criterion for prediction-only studies.
- Quentin Cappart, Didier Chételat, Elias B Khalil, Andrea Lodi, Christopher Morris, and Petar Veličković. Combinatorial optimization and reasoning with graph neural networks. *Journal of Machine Learning Research*, 24(130):1–61, 2023. URL <https://www.jmlr.org/papers/volume24/21-0449/21-0449.pdf>. **Excluded:** Although the paper discusses several neural-symbolic approaches to propositional satisfiability and solver guidance, it is a conceptual

review and does not introduce or evaluate an original PL reasoning framework. The relevant primary studies are considered separately.

David "davidad" Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua Tenenbaum. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems, 2024. URL <https://arxiv.org/abs/2405.06624>.

Ioannis Hatzilygeroudis and Jim Prentzas. Neurules: improving the performance of symbolic rules. *International Journal on Artificial Intelligence Tools*, 9(01):113–130, 2000. URL <https://www.worldscientific.com/doi/abs/10.1142/S0218213000000094>. **Excluded:** The paper presents neurules, hybrid rules integrating production rules with an adaline unit, to improve rule-base size and inference efficiency in expert systems. It does not explicitly perform SAT solving, satisfiability decision, satisfiability witness generation, UNSAT refutation, propositional entailment, resolution, proof search, proof generation, or proof validation for PL.

Steffen Hölldobler and Yvonne Kalinke. Approximating the semantics of logic programs by recurrent neural networks. *Applied Intelligence*, 11(1):45–58, 1999. URL <https://link.springer.com/content/pdf/10.1023/A:1008376514077.pdf>. **Excluded:** The paper’s new neural-network construction approximates the semantics of first-order acyclic logic programs; its discussion of propositional logic programs only recalls a construction established in earlier work.

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943. URL <https://jontalle.web.engr.illinois.edu/uploads/410-NS.F22/McCulloch-Pitts-1943-neural-networks-ocr.pdf>.

Ling Sun, David Gerault, Adrien Benamira, and Thomas Peyrin. NeuroGIFT: Using a Machine Learning Based Sat Solver for Cryptanalysis. In Shlomi Dolev, Vladimir Kolesnikov, Sachin Lodha, and Gera Weiss, editors, *Cyber Security Cryptography and Machine Learning*, volume 12161 of *Lecture Notes in Computer Science*, pages 62–84, Cham, 2020. Springer. doi: 10.1007/978-3-030-49785-9_5. URL https://doi.org/10.1007/978-3-030-49785-9_5. **Excluded:** NeuroGIFT predicts satisfiability for SAT problems arising from differential cryptanalysis of GIFT, but the reported results establish classification accuracy without demonstrating PL reasoning or proving beyond prediction. The authors could not extract satisfying assignments from NeuroGIFT-V1 or NeuroGIFT-V2 (Sections 4.1–4.2, author-hosted manuscript, pp. 15–17). They leave confirmation of actual SAT solving and neural guidance of classical solvers to future work (p. 19).

Joseph Townsend, Thomas Chaton, and Joao M Monteiro. Extracting relational explanations from deep neural networks: A survey from a neural-symbolic perspective. *IEEE transactions on neural networks and learning systems*, 31(9):3456–3470, 2019. URL <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8889997>.

Athanasios Tsakonas, Vasilios Aggelis, Ioannis Karkazis, and Georgios Dounias. An evolutionary system for neural logic networks using genetic programming and indirect encoding. *Journal of Applied Logic*, 2(3):349–379, 2004. ISSN 1570-8683. doi: <https://doi.org/10.1016/j.jal.2004.03.005>. URL <https://www.sciencedirect.com/science/article/pii/S1570868304000199>. **Excluded:** Accessible abstract/metadata describe an evolutionary system that uses genetic programming and indirect encoding to produce neural logic networks for classification and interpretable expert-rule extraction. The accessible sources do not show an explicit PL SAT, satisfiability decision, satisfiability witness generation, UNSAT refutation, entailment, reso-

lution, proof-search, proof-generation, or proof-validation task. t does not explicitly show PL formulas used for SAT, entailment, proof, or refutation. Genetic programming guides the construction/topology of neural logic networks. This alignment is for classification / knowledge extraction, not an explicit PL reasoning or proving algorithm.