Article

Not peer-reviewed version

# Responsible Reasoning - a Systematic Review

Jason Pittman *, Lindsay Eddy, Kyle Wiseman

Posted Date: 15 October 2024

doi: 10.20944/preprints202410.0985.v1

Keywords: Responsible AI; Neurosymbolic AI; Controls; Guardrails; Artificial Intelligence

 doi:10.20944/preprints202410.0985.v1



*Article*

# Responsible Reasoning - A Systematic Review

**Jason M. Pittman *, Lindsay Eddy and Kyle Wiseman**

Booz Allen Hamilton
* Correspondence: pittman_jason@bah.com

**Abstract:** The integration of responsible artificial intelligence (RAI) principles with emerging neurosymbolic AI (NSAI) systems is crucial for the development of fair, explainable, and trustworthy AI technologies. This paper presents a systematic review exploring the convergence of RAI and NSAI, analyzing current research to assess how RAI principles such as explainability, bias, robustness, transparency, and privacy have been applied to NSAI. This work employed a systematic literature review to synthesize findings from a sample of papers demonstrating RAI principle implementations. Our analysis reveals two main trends: significant research demonstrates the application of NSAI to enhance RAI principles in other AI systems, while limited work directly applies RAI principles to NSAI architectures. Key challenges include the lack of established frameworks for implementing RAI within NSAI systems and the complexities inherent in merging neural and symbolic reasoning methods. This review highlights open research gaps and suggests pathways for future work, emphasizing the need for robust RAI frameworks tailored to NSAI systems.

**Keywords:** Responsible AI; Neurosymbolic AI; Controls; Guardrails; Artificial Intelligence

---

## I. Introduction

Rapid advancements in artificial intelligence (AI) have been driven by the successes of deep learning techniques Garcez *et al.* (2019). Deep learning has demonstrated improved accuracy and performance compared to pre-existing types of AI systems LeCun *et al.* (2015); Wan *et al.* (2024). Other benefits of deep learning include learning from unstructured data Krizhevsky *et al.* (2012) and independence from intermediate feature engineering Hinton *et al.* (2006). However, deep learning has two serious limitations. The first is opacity of decision-making. Meaning, deep learning systems have difficulty in evidencing how and why a given decision was made. Further, a second limitation is an inability to incorporate structured knowledge into a decision while facing high levels of uncertainty.

Neurosymbolic AI seeks to bridge the gap between the data-driven capabilities of (deep) neural networks and the reasoning power of symbolic systems. By integrating these two paradigms, neurosymbolic AI aims to create more robust, explainable, and efficient AI systems. More specifically, NSAI systems, according to Hitzler et al. Hitzler and Sarker (2022), combine neural networks with predicate logic. Such systems are capable of reasoning about complex problems in environments with high levels of uncertainty Wan *et al.* (2024). NSAI is expected to have a broad yet significant impact on the future of AI systems because of the ability to handle such circumstances.

Fields such as healthcare, education, and robotics are experiencing benefits from NSAI-based innovations Campagner and Cabitza (2020); Inala (2022); Wagner and d'Avlia Garcez (2024). Yet, any field relying on reasoning over large datasets within a set of rules or facts will see benefits from NSAI over traditional AI systems such as deep learning. Furthermore, NSAI has potential to establish generalized explainability and trustworthiness in other AI systems Wan *et al.* (2024).

This last notion integrates NSAI with another burgeoning field- Responsible AI (RAI). RAI emphasizes the need for AI systems of any type to be fair, trustworthy, and aligned with societal values Mitchell *et al.* (2019). The field has established a consistent set of principles. Responsible AI controls, such as explainability, fairness, and robustness, are crucial to addressing these challenges and ensuring the development of trustworthy AI systems. Further, there has been demonstrable success Speith (2022); Hort *et al.* (2024); Upreti *et al.* (2024) related to applying RAI principles to traditional AI systems such as neural networks, classifiers, and so forth. Additionally, foundational work is

 doi:10.20944/preprints202410.0985.v1



underway exploring how RAI principles may be applied to generative AI Kim *et al.* (2024). However, challenges related to explainability, trustworthiness, fairness, robustness and safety, as well as privacy, applied to NSAI remain unresolved Wan *et al.* (2024); Hitzler and Sarker (2022); Hamilton *et al.* (2022); Delong *et al.* (2023).

Generally speaking, ensuring AI systems operate ethically and responsibly is a critical initiative for researchers and practitioners Cheng *et al.* (2021). AI systems of any type can develop divergent qualities or behaviors when responsible AI principles are absent. For this reason, the purpose of this work was to assess the state of knowledge in regard to the convergence of RAI principles and NSAI. More specifically, this study sought to uncover which, if any, research demonstrated technical implementations of RAI principles in NSAI systems.

The rest of this paper is organized as follows. The next section presents a conceptual framework through related work. The framework consists of definitions for RAI and NSAI, significance of each field, as well as open challenges in each. Then, the method employed to fulfill the purpose of this study is discussed.

## II. Related Work

The related work supporting this systematic review consists of two converging literatures: NSAI and RAI. The aim of this section is to impart a sufficient understanding of definitions, significance of NSAI and RAI, as well as highlighting key open challenges in each field.

### *A. Neurosymbolic AI*

While deep learning has achieved remarkable success in various fields, it has limitations, such as the lack of interpretability and the requirement for large amounts of labeled data. NSAI addresses these issues by integrating symbolic reasoning, which can leverage existing knowledge and provide explanations for the AI's decisions. In addition, the inclusion of probabilistic approaches in NSAI helps in dealing with uncertainty and improving the robustness of AI systems. This is particularly valuable in real-world applications where data can be noisy or incomplete Wan *et al.* (2024).

Neurosymbolic AI combines machine learning methods based on artificial neural networks (such as deep learning) with symbolic approaches to computing and AI, such as those found in knowledge representation and reasoning Hitzler and Sarker (2022); Hamilton *et al.* (2022); Delong *et al.* (2023). Early works like those by Besold et al. Besold *et al.* (2017) laid the groundwork by exploring basic integration of neural and symbolic methods. Such early work focused on improving the interpretability and reasoning capabilities of neural networks. From there, three areas of significance emerged for NSAI. These areas are offsetting of inherent limitations in deep learning, improved handling of uncertainty, and a potential step towards artificial general intelligence.

Fundamentally, NSAI differs from traditional AI systems such as neural networks, classifiers, and regression models by because of the symbolic reasoning layer. Traditional AI models like neural networks excel at learning from large datasets but often struggle with interpretability and reasoning. Symbolic AI, on the other hand, excels in logical reasoning Dong *et al.* (2019). Hence, NSAI fills in gaps in traditional AI systems. Doing so enables systems to perform data-driven predictions while also applying high-level symbolic reasoning Garcez *et al.* (2019).

This hybrid approach allows for greater flexibility in solving complex problems, offering advantages in areas where traditional models may fall short, such as generalization, intepretability, and explainability (Sarker et al., 2021). Moreover, NSAI is designed to handle complex reasoning tasks more efficiently, mimicking human-like cognitive processes by blending the interpretability of symbolic AI with the adaptability of neural networks (Besold et al., 2017).

One of the long-term goals of NSAI is to contribute to the development of human-level AI, which combines the learning capabilities of neural networks with the logical reasoning abilities of symbolic systems. This interdisciplinary approach, according to Wan Wan *et al.* (2024) is essential for creating

 doi:10.20944/preprints202410.0985.v1



AI systems that can perform complex cognitive tasks and exhibit human-like understanding and problem-solving skills.

1) Challenges and Opportunities

Merging neural networks with symbolic reasoning is inherently challenging. Neural networks excel at pattern recognition but lack interpretability, while symbolic systems are interpretable but struggle with ambiguity and noise d'Avila Garcez and Lamb (2020). Thus, achieving a seamless integration without compromising the strengths of each approach remains difficult. Furthermore, NSAI is suitably more complex than previous AI systems. Dong et al. Dong *et al.* (2019) found incorporating advanced logical reasoning into neural architectures engenders significant computational overhead. Indeed, integrating symbolic reasoning can sometimes lead to decreased performance or increased complexity in neural models Li and Srikumar (2019).

Given time, it is reasonable to suspect NSAI researchers will overcome most or all of these challenges. However, there are related and adjacent RAI challenges which are not so easily addressed. For instance, the complexity and performance challenges can have negative impacts on explainability Li and Srikumar (2019), interpretability Cunnington *et al.* (2022), as well as a variety of robustness and safety parameters Besold *et al.* (2017). Furthermore, the literature Garcez *et al.* (2019); Eskov *et al.* (2021); Shakarian and Simari (2022) suggests there is a lack widely accepted framework or set of best practices for developing NSAI systems.

*B. Responsible AI*

AI is a mainstream technology and highly embedded in culture. Much less common is how ethical and responsible AI can be achieved although, according to the literature Jobin *et al.* (2019); Arrieta *et al.* (2020); Mehrabi *et al.* (2021), there is increasing demand for such. Definitionally, responsible AI ensures AI systems are developed and deployed in ways that are ethical Floridi *et al.* (2018); Mittelstadt *et al.* (2016). Ethical, in this context, implies principles such as fairness, transparency, privacy, security, and trustworthiness. The idea is an AI system can be considered responsible when the set of relevant principles are present. Of course, to be present implies some form of evaluation or assessment.

To that end, ethical principles have gone through rapid theoretical and practical expansion over the past decade. In this short time, researchers have developed robust technical frameworks to measure and evaluate these principles. Two prominent examples are the Microsoft Responsible Toolbox and the IBM AI 360 Toolkit. AI practitioners can use these frameworks to evaluate models. Yet, researchers Radclyffe *et al.* (2023); Lu *et al.* (2024) suggest RAI is one of the most critical challenges present in the broader AI field of study.

Culturally, the rapid expansion has been driven by notable examples of harm resulting from a lack of responsible AI. Such examples include discriminatory sentencing and parole decisions in the U.S. justice system Angwin *et al.* (2022) and Amazon's recruitment tool becoming biased against women Dastin (2022). Another part of the expansion is increasing legal and regulatory requirements such as U.S. President Biden's Executive Order and the EU's AI Act Wörsdörfer (2023).

Meanwhile, the literature Khan *et al.* (2022); Alzubaidi *et al.* (2023) has coalesced around five specific RAI principles: explainability, bias or fairness, robustness or safety, transparency or interpretability, and privacy. Additional principles, such as explicability Prem (2023) and accountability Liu *et al.* (2022), have been studied but ultimately fall within the scope of one or more of the five specific principles. Consequently, industry (IBM, Microsoft, US Department of Defense) has settled on explainability, bias, robustness, interpretability, and privacy for practical RAI implementation.

1) Explainability

To that end, explainability is understood to be an AI system's ability to *explain* its behaviors and outcomes Arrieta *et al.* (2020); Hoffman *et al.* (2018). The field views behavior or outcome as proxies for decision-making. The principle seeks to clarify how AI systems reach specific conclusions, making

 doi:10.20944/preprints202410.0985.v1



them understandable to human operators. Notably, explainability is tightly coupled to the technical interpretability or transparency of AI system inner workings.

2) Bias or Fairness

Biased AI systems exhibit skewed outputs based on prejudiced inputs Hort *et al.* (2024). Often, *bias* is understood as affecting individuals based on demographics Mehrabi *et al.* (2021). This is true. However, AI system bias also may result from preferential data ingestion from one sensor in an array or unequal, non-demographic feature weighting Blasch *et al.* (2021). *Fairness*, then, as the companion technical principle aims to prevent bias by ensuring equitable treatment and outcomes across different groups (persons or systems). Such can apply to data, algorithms, or outputs.

3) Robustness or Safety

When an AI system maintains reliable performance across a wide range of conditions, including noisy or adversarial inputs, distribution shifts, and unforeseen changes in the environment, the literature deems such to be *robust* Hendrycks and Gimpel (2016); Goodfellow *et al.* (2014). Closely related, *safety* ensures AI systems behave in a predictable, controlled, and secure manner, even in the presence of unexpected challenges or adversarial manipulations Raji and Dobbe (2023). Together, these concepts assure AI systems from errors, vulnerabilities, and harmful outcomes.

4) Interpretability or Transparency

Interpretability refers to the extent to which human operators can comprehend and reason about the explanations an AI system provides Doshi-Velez and Kim (2017); Gilpin *et al.* (2018). The principle renders the internal logic *transparent* such that operators understand how input data is transformed into outputs. Significantly, detailed knowledge of the model's algorithmic structure is not, and cannot, be required. Then, in combination with explainability, operators can access the completely pipeline of AI system decision-making.

5) Privacy

The RAI principle *privacy* protects sensitive information from misuse, exposure, or unauthorized accessed Sweeney (2002). In this way, AI system privacy strategies minimize risk of data breaches, unauthorized surveillance, and re-identification of individuals Shokri and Shmatikov (2015). Privacy is differentiated, in simple terms, from robustness and safety because the latter works to stop something from happening whereas privacy reveals when something has happened. The two function best when paired similar to explainability and interpretability. Unique to the five RAI principles, AI system privacy offers technical mechanisms to comply with international governance policies (e.g., GDPR).

6) Challenges and Opportunities

Despite the stated need for RAI and availability of broad technical frameworks, the field has a variety of open research challenges. Such is observable given how the design and implementation of responsible AI principles continues to appear as ideas for future work throughout the literature Whittlestone and Clark (2021); Fjeld *et al.* (2020). Specific examples include, but are not limited to, developing trustworthy models that are transparent and interpretable is problematic Lundberg and Lee (2016). Protecting AI systems from adversarial attack Goodfellow *et al.* (2014); Papernot *et al.* (2016) is also an open challenge. Moreover, because AI systems are dependent upon data, ensuring privacy of personal or otherwise sensitive data is a nontrivial aspect of ongoing research Abadi *et al.* (2016); Wei *et al.* (2021).

Furthermore, two gaps become obvious in the literature when inferring whitespace between frontier innovations in AI and nascent responsible AI research. Foremost, there is little or no guidance for practitioners. While researchers have presented technical responsible AI implementations for traditional AI systems, there is nothing to connect concept to discrete application. Moreover, the cutting edge of AI research (i.e., NSAI) seems to have expanded rapidly beyond the RAI horizon.

 doi:10.20944/preprints202410.0985.v1



Thus, there should be little surprise that similar RAI challenges surround NSAI as is true for traditional AI systems, at least in the neural network layer. For example, Hitzler et al. Hitzler and Sarker (2022) suggested fairness can be assured through transparency and explainability. Yet, Wan et al. Wan *et al.* (2024) articulated a need for enhanced explainability and trustworthiness in NSAI systems. The contradiction causes confusion and leaves a significant gap in the literature. Accordingly, it is not clear in the literature how one would go about implementing RAI principles in NSAI. This systematic review aims to address the lack of clarity.

## III. Method

This work was motivated by a single research question: what RAI principles have demonstrated implementations for NSAI systems? Our aim with such a question was twofold. On one hand, this question drove a synthesis of what RAI principles have demonstrated application to NSAI systems. On the other hand, by proxy, this question would reveal gaps where RAI principles have not yet been applied to NSAI systems.

A systematic literature review design facilitated collecting and analyzing relevant research to answer the research question. As part of the review, multiple online public databases were queried such as Google Scholar, arXiv, IEEEXplore, ACM Digital Library, and DBLP Computer Science Bibliography. Date ranges during the literature searches were not restricted. Further, duplicate papers were removed from the collection before proceeding. A manual inspection of the NSAI related papers was performed and each paper was evaluated according to the inclusion-exclusion criteria.

### *A. Search Strategies*

Our search strategy consisted of iterative queries using a set of RAI principles (*explainability, bias or fairness, robustness or safety, interpretability or transparency, privacy*) and a set of AI types (*neurosymbolic AI or NSAI, symblolic AI, AI,and machine learning or ML*). Boolean AND/OR operators were used to combine keywords from each set into rational search strings. Two examples of rational search strings would be "explainability AND (neurosymbolic AI OR NSAI)" and "explainability AND AI".

Broad searches were intentionally employed to begin with to minimize the chance of missing even tangentially related papers. Further, one search strategy included general AI and machine learning as terms. Doing so was a means to paint a contrast. Table 1 summarizes the literature discovery. The *count* is the total articles returned from the search.

**Table 1.** Literature search strings with count of discovered papers.

| Search String | Count |
|---|---|
| explainability AND (neurosymbolic AI OR symbolic AI OR NSAI) | 3,040 |
| explainability AND ((machine learning OR ML) OR AI) | 157,000 |
| (bias OR fairness) AND (neurosymbolic AI OR symbolic AI OR NSAI) | 3,260 |
| (bias OR fairness) AND ((machine learning OR ML) OR AI) | 3,280,000 |
| (robustness OR safety) AND (neurosymbolic AI OR symbolic AI OR NSAI) | 3,310 |
| (robustness OR safety) AND ((machine learning OR ML) OR AI) | 4,490,000 |
| (interpretability or transparency) AND (neurosymbolic AI OR symbolic AI OR NSAI) | 1,420 |
| (interpretability or transparency) AND ((machine learning OR ML) OR AI) | 72,200 |
| privacy AND (neurosymbolic AI OR symbolic AI OR NSAI) | 2,540 |
| privacy AND ((machine learning OR ML) OR AI) | 5,140,000 |

*Note*: Counts are minimum estimates based on publicly available search results. More research may be available beyond our searches.

### *B. Inclusion and Exclusion Criteria*

Full-text NSAI papers with a publicly accessible document were included. Whereas, discovered papers with only a public abstract were excluded. Further, papers containing a demonstrated technical RAI principle implementation- inclusive of journal papers, conference papers, theses, and dissertations- were included. Demonstrated application in the context of the literature equated to either sufficient technical details to construct an implementation or a sample implementation available in pseudocode

 doi:10.20944/preprints202410.0985.v1



or source code. Literature not containing one or the other were not included in this systematic review. As well, papers demonstrating policy, governance, or otherwise non-technical expressions of RAI principles were excluded.

The final literature sample after applying the inclusion-exclusion criteria consisted of 25 NSAI related papers and and 974 general AI or ML papers in total. Down-selection outcomes were tracked by search string category (Table 2). The *count* represents the resulting total after inclusion and exclusion criteria were applied. Criteria were continually applied to the total search results until either reaching duplicate saturation or exhausting the dataset.

**Table 2.** Subsets of discovered papers selected for analysis.

| Search String | Count |
|---|---|
| explainability AND (neurosymbolic AI OR symbolic AI OR NSAI) | 11 |
| explainability AND ((machine learning OR ML) OR AI) | 200[1] |
| (bias OR fairness) AND (neurosymbolic AI OR symbolic AI OR NSAI) | 3 |
| (bias OR fairness) AND ((machine learning OR ML) OR AI) | 200[1] |
| (robustness OR safety) AND (neurosymbolic AI OR symbolic AI OR NSAI) | 7 |
| (robustness OR safety) AND ((machine learning OR ML) OR AI) | 192 |
| (interpretability or transparency) AND (neurosymbolic AI OR symbolic AI OR NSAI) | 2 |
| (interpretability or transparency) AND ((machine learning OR ML) OR AI) | 182 |
| privacy AND (neurosymbolic AI OR symbolic AI OR NSAI) | 1 |
| privacy AND ((machine learning OR ML) OR AI) | 200[1] |

*Note*: Counts are minimum estimates based on publicly available search results. More research may be available beyond our searches. [1] Collection of ML / AI counts stopped after 20 pages of search results (at 10 results per page). [2] Total NSAI literature count is 24 and did not include *trusthworthy* because the principle was emergent during analysis. See section IV.

### C. Information Extraction

Information was extracted from the collected papers by inspecting the full-text for technical responsible AI principle implementation details. Then, papers were sorted into categories using literature dimensions such as published date, keywords, RAI principle(s), and whether each principle applied to NSAI or was NSAI applying the principle to another AI system. For completeness, citation metadata was also extracted so as to map potential relationships between research meeting our inclusion criteria.

## IV. Findings

Recall this study set out to determine what RAI principles have demonstrated implementations for NSAI systems. To accomplish this, NSAI literature spanning four years was analyzed. The oldest study was published in 2020 while the most recent appeared in 2024 (Figure 1). There were three papers published 2020 with a steady upwards trend reaching eight papers in 2023. The year 2024 had six papers published with four months remaining.

Sources for the papers varied between seven entities. The most frequent source were conference proceedings. The next highest frequency of papers came from arXiv preprints. One thesis and two dissertations contributed to the findings. Professional society journals supplied two papers, one each from IEEE and ACM. Finally, the remaining papers came from a diverse array of journals.

There was only one instance of repetition of primary author across the papers Wagner and d'Avlia Garcez (2024); Wagner and d'Avila Garcez (2021). Additionally, one paper appeared in two different groupings Amado *et al.* (2023). There were no citation connections between the papers analyzed. Meaning, no given paper cited another paper in the dataset.

As an aside, one may notice the introduction of a sixth RAI principle- *trustworthy* or *trustworthiness*. While precedent exists for encapsulating RAI principles under the label of trustworthy, this study found NSAI research treating trustworthiness distinct from other principles (e.g., explainability *and trustworthiness*. Therefore, such papers were analyzed separately.

 doi:10.20944/preprints202410.0985.v1



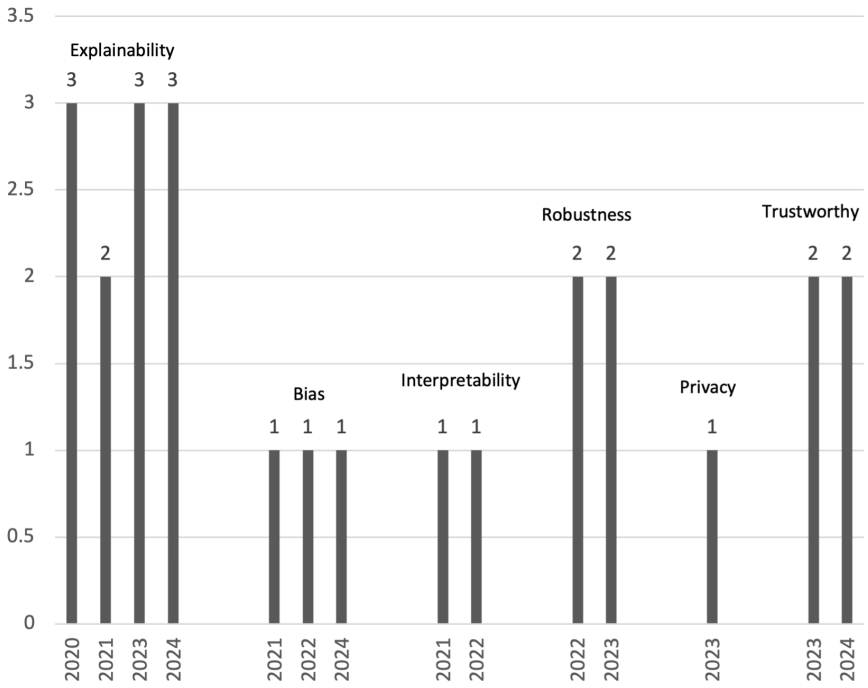


**Figure 1.** Number of analyzed papers grouped year of publication and by RAI principle.

In total, 25 papers were analyzed (Figure 2). Forty-four percent of the papers demonstrated the RAI principle of explainability. Robustness as a principle represented the next largest cluster at 16%. Thereafter, the collected papers demonstrated bias, interpretability, and privacy at 12%, 8%, and 4% respectively. Trustworthiness, the emergent principle, accounted for 16% of the analyzed literature.

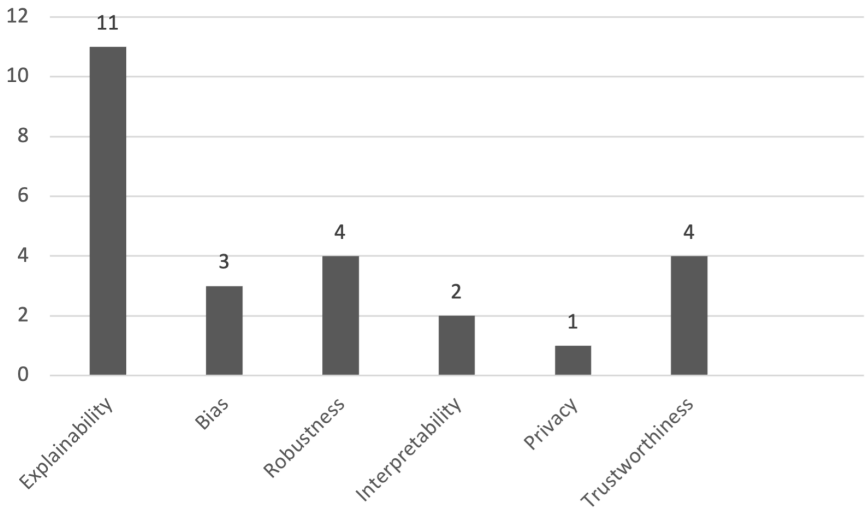


**Figure 2.** Number of analyzed papers grouped by RAI principle.

Furthermore, the literature was split between two directions. The majority- 84%- of the literature exhibited the application of NSAI to other AI systems for the purposes of implementing a RAI principle (Table 3. The other direction constituted a RAI principle applied to a NSAI system (Table 4. Such work comprised 16% of the analyzed research. These percentages were inclusive of the trustworthiness principle.

 doi:10.20944/preprints202410.0985.v1



**Table 3.** Literature demonstrating NSAI applying RAI principles to AI systems.

| Principle | Reference | Year | Technique |
|---|---|---|---|
| Explainability | Pisano *et al.* (2020) | 2020 | prototype integrating symbolic logic into sub-symbolic systems |
| | Oltramari *et al.* (2020) | 2020 | hybrid system combining data-driven perception with logical reasoning |
| | Campagner and Cabitza (2020) | 2020 | proof of concept using Logic Tensor Networks and rule-based systems |
| | Venugopal *et al.* (2021) | 2021 | framework providing uncertainty estimates for its predictions |
| | Himmelhuber *et al.* (2021) | 2021 | a fidelity metric using graph neural networks and symbolic logic |
| | d'Avila Garcez and Lamb (2020) | 2023 | fidelity and soundness measures based on distributed and local symbols |
| | Bellucci (2023) | 2023 | ontology-based image classifier using a structured knowledge base |
| | | | data visualization, feature importance analysis, and partial dependence plots (PDPs) |
| | Dwivedi *et al.* (2023) | 2023 | permutation feature importance and SHAP values |
| | | | counterfactuals and contrastive explanations |
| | | | post-hoc interpretations of model predictions with LIME |
| | | | software libraries such as Skater or AIX360 |
| | Mileo | 2024 | framework to integrate human feedback, causal reasoning, and knowledge injection |
| | Wagner and d'Avlia Garcez (2024) | 2024 | framework for a logic-based querying system |
| | Thota and Arora (2024) | 2024 | human-centric interactive interface with knowledge graph integration |
| Bias | Wagner and d'Avila Garcez (2021) | 2021 | framework using SHAP measure with demographic parity and disparate impact metrics |
| | Xie *et al.* (2022) | 2022 | Neuro-Symbolic Assertion Language to formalize fairness properties enforced with specification networks |
| | Padalkar *et al.* (2024) | 2024 | NeSyBiCor framework using Answer Set Programming with semantic similarity measure |
| Interpretability | Hooshyar and Yang (2021) | 2021 | framework focused on knowledge representation, symbolic constraints, and knowledge extraction |
| | Bennetot *et al.* (2022) | 2022 | Greybox XAI framework with deep neural network (DNN) building Explainable Latent Space |
| Robustness | Smirnova *et al.* (2022) | 2022 | Nessy system uses expectation regularization and data sampling |
| | Inala (2022) | 2022 | use of state machines and neurosymbolic transformers for formal verification |
| | Amado *et al.* (2023) | 2023 | Predictive Plan Recognition (PPR) framework removes noise and gaps |
| Privacy | Piplai *et al.* (2023) | 2023 | framework combining differential privacy, secure multi-party computation, and synthetic data generation |
| Trustworthiness | Zeng (2024) | 2024 | framework integrating differentiable learning with graph neural network rewiring |

On one hand, the analyzed NSAI for RAI research in Table 3 demonstrated three types of techniques: prototypes, measures, and frameworks. Across the 21 papers, eight constituted some kind of prototype (prototype, system, or proof of concept). Measures appeared four times. Frameworks appeared most frequently with nine occurrences. On the other hand, the papers showing application of RAI principles to NSAI systems in Table 4 revealed two of the three techniques from the previous direction. Prototypes and frameworks were evenly distributed with two each. Measures were not represented.

 doi:10.20944/preprints202410.0985.v1



**Table 4.** Literature demonstrating RAI principles applied to NSAI to AI systems.

| Principle | Reference | Year | Technique |
|---|---|---|---|
| Trustworthiness | Agiollo and Omicini (2023) | 2023 | NeSy system combining various RAI principles |
| | Kosasih *et al.* (2023) | 2023 | hybrid architecture using neural network data-driven learning and the symbolic rules |
| | Gaur and Sheth (2024) | 2024 | CREST framework combining procedural and graph-based knowledge with neural network capabilities |
| Robustness | Amado *et al.* (2023) | 2023 | Predictive Plan Recognition (PPR) framework removes noise and gaps |

## V. Conclusion

Deep learning is at the core of modern AI mainstream popularity Garcez *et al.* (2019). AI systems such as ChatGPT are possible because of the enhanced capabilities of deep learning architecture. Yet, deep learning decisions are opaque and the systems falter when facing high uncertainty. NSAI aims to address these gaps by integrating neural networks with symbolic computing Hitzler and Sarker (2022). In short, NSAI adds a reasoning capability which is transparent and can handle high degrees of uncertainty during decision making.

While the capabilities of NSAI address the gaps in deep learning, all AI systems are subject to ethical and responsible controls. Once implemented RAI principles render AI systems of any type explainable, unbiased, interpretable, robust, and trustworthy. Traditional AI systems such as classifiers, regression models, and clustering systems have a rich literature available in this area. In fact, the research demonstrates a plethora of RAI techniques across all RAI principles (Table 2). Until this work, the depth and breadth of RAI for NSAI was unknown. Thus, the purpose of this work was to assess the state of knowledge in regard to the convergence of RAI principles and NSAI.

A systematic review design facilitated the collection and analysis of pertinent research. The initial search uncovered 13,570 papers. After applying inclusion-exclusion criteria, the sample consisted of 25 papers. From this collection, the systematic review revealed two overarching features of the converged RAI and NSAI literature. First, substantial research exists demonstrating the application of NSAI for RAI principles. Such included the discovery of an emergent principle in trustworthiness. Second, much less research exists demonstrating application of RAI principles to a NSAI system. The reasoning behind these features might be best understood in three parts.

### *A. Inferences*

Recall explainability comprised a significant quantity of existing NSAI for RAI research. One may infer the focus on explainability, at least in part, has been inherited from push for explainability in traditional AI systems. Deep learning especially is limited because of opaqueness but so are the various other traditional AI systems. NSAI innately addresses explainability because of its reasoning capability. Therefore, explainability representing a significant portion of NSAI for RAI research is unsurprising.

The same rationale hints at one of the three reasons for the discovered features of the sample. That is, NSAI as a type of AI, is adept at applying RAI principles as a consequence of being able to reason. Indeed, one can observe the necessity of reasoning in both the theoretical and applied RAI literature Selbst *et al.* (2019); Christoph (2020); Müller (2020).

The second part of our rational has two sub-parts. On one hand, we observed an extensive literature for RAI principle implementations in traditional AI systems. On the other hand, the use of the term *trustworthy* in combination with explainability within NSAI for RAI literature was somewhat surprising. The associated research makes clear the term trustworthy encompasses multiple RAI principles Di Maio (2020). Yet, the treatment of explainability apart from other RAI principles is a curious matter.

 doi:10.20944/preprints202410.0985.v1



Explainability separate from trustworthiness is a curious matter insofar as NSAI research presupposes if explainability is correctly implemented, then the other RAI principles (being subordinate) must likewise be present. Stated differently, the other RAI principles are implicitly present by virtue of explainabilty being present. A further thought might be explainability is not implementable as a solo principle in NSAI.

Lastly, the third part is implicit in the power of NSAI for RAI and connects back to the first part. NSAI can apply RAI principles to itself, specifically the neural network layer. Such would also be true for any multi-modal AI architecture embedded below the reasoning layer. Then, because NSAI is innately explainable, either a human-in or human-on the loop can reason about the ethical and responsible nature of a NSAI system's outputs.

### *B. Limitations*

The above tripartite rationale has limitations as does this study, however. It is possible the entire presupposition is incorrect. The prevalence of NSAI for RAI research compared to RAI for NSAI could be skewed because of flaws in our systematic review execution. Whereas, even if our systematic review execution was sound, it is possible research exists outside of the indices searched. If true, this changes the distribution of collected papers. Moreover, the dearth of RAI for NSAI research might reflect deep challenges, even impossibilities, in RAI principle implementations within an NSAI system.

### *C. Future Work*

Overall, tremendous opportunity exists at the intersection of RAI and NSAI. There exists opportunity both in applying RAI principles to NSAI as well as using NSAI to apply RAI principles. As well, the stated limitations are addressable in future work. To that end, there are three specific areas of potential study as follows.

A preeminent area for study is the implementation of RAI principles demonstrated in traditional AI literature to NSAI architectures. A sequence consisting of reproduction or replication of traditional AI research for each principle, constructive work in porting each principle to NSAI architectures may be beneficial. Such could be followed up by work investigating the impact of all principles within a NSAI architecture.

Based on the outcomes from any future work in the prior category, studying impact of RAI principle implementation on NSAI explainability may have significance. To the extent NSAI explainability is propositionally related to the other RAI principles, knowing whether all, some, just one, or none of the other principles is sufficient for explainability. Moreover, such work can investigate whether one specific technique of an individual RAI principle from traditional AI research is more or less suited for NSAI. Reproduction or replication study, as well as constructive work, may not be necessary for this line of inquiry.

Finally, future work might investigate the extent to which an NSAI system may fulfill the human-in or human-on the loop role when evaluating the ethical and responsible state of a given NSAI system. Framed another way, there is an opportunity to explore the use of NSAI against NSAI for RAI principle implementation. Such work may investigate explainable NSAI system A evaluating the explainability associated with NSAI system B. Also, work could look into NSAI system A reasoning about the other individual RAI principles in NSAI system B.

## References

Garcez, A.d.; Gori, M.; Lamb, L.C.; Serafini, L.; Spranger, M.; Tran, S.N. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *arXiv preprint arXiv:1905.06088* **2019**.

LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436–444.

Wan, Z.; Liu, C.K.; Yang, H.; Li, C.; You, H.; Fu, Y.; Wan, C.; Krishna, T.; Lin, Y.; Raychowdhury, A. Towards cognitive ai systems: a survey and prospective on neuro-symbolic ai. *arXiv preprint arXiv:2401.01040* **2024**.

 doi:10.20944/preprints202410.0985.v1




Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **2012**, *25*.

Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural computation* **2006**, *18*, 1527–1554.

Hitzler, P.; Sarker, M.K. *Neuro-symbolic artificial intelligence: The state of the art*; IOS press, 2022.

Campagner, A.; Cabitza, F. Back to the feature: A neural-symbolic perspective on explainable AI. Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4. Springer, 2020, pp. 39–55.

Inala, J.P. Neurosymbolic Learning for Robust and Reliable Intelligent Systems. PhD thesis, Massachusetts Institute of Technology, 2022.

Wagner, B.J.; d'Avlia Garcez, A. A neurosymbolic approach to AI alignment. *Neurosymbolic Artificial Intelligence* **2024**, pp. 1–12.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model cards for model reporting. Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 220–229.

Speith, T. A review of taxonomies of explainable artificial intelligence (XAI) methods. Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, 2022, pp. 2239–2250.

Hort, M.; Chen, Z.; Zhang, J.M.; Harman, M.; Sarro, F. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing* **2024**, *1*, 1–52.

Upreti, R.; Lind, P.G.; Elmokashfi, A.; Yazidi, A. Trustworthy machine learning in the context of security and privacy. *International Journal of Information Security* **2024**, *23*, 2287–2314.

Kim, S.; Cho, J.Y.; Lee, B.G. An Exploratory Study on the Trustworthiness Analysis of Generative AI. *Journal of Internet Computing and Services* **2024**, *25*, 79–90.

Hamilton, K.; Nayak, A.; Božić, B.; Longo, L. Is neuro-symbolic ai meeting its promises in natural language processing? a structured review. *Semantic Web* **2022**, pp. 1–42.

Delong, L.N.; Mir, R.F.; Whyte, M.; Ji, Z.; Fleuriot, J.D. Neurosymbolic ai for reasoning on graph structures: A survey. *arXiv preprint arXiv:2302.07200* **2023**, *2*.

Cheng, L.; Varshney, K.R.; Liu, H. Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research* **2021**, *71*, 1137–1181.

Besold, T.R.; Garcez, A.d.; Stenning, K.; van der Torre, L.; van Lambalgen, M. Reasoning in non-probabilistic uncertainty: Logic programming and neural-symbolic computing as examples. *Minds and Machines* **2017**, *27*, 37–77.

Dong, H.; Mao, J.; Lin, T.; Wang, C.; Li, L.; Zhou, D. Neural logic machines. *arXiv preprint arXiv:1904.11694* **2019**.

d'Avila Garcez, A.; Lamb, L.C. Neurosymbolic AI: The 3rd wave. *arXiv e-prints* **2020**, pp. arXiv–2012.

Li, T.; Srikumar, V. Augmenting neural networks with first-order logic. *arXiv preprint arXiv:1906.06298* **2019**.

Cunnington, D.; Law, M.; Lobo, J.; Russo, A. Neuro-symbolic learning of answer set programs from raw data. *arXiv preprint arXiv:2205.12735* **2022**.

Eskov, V.M.; Filatov, M.A.; Gazya, G.; Stratan, N. Artificial intellect with artificial neural networks. *Russian Journal of Cybernetics* **2021**, *2*, 44–52.

Shakarian, P.; Simari, G.I. Extensions to generalized annotated logic and an equivalent neural architecture. 2022 Fourth International Conference on Transdisciplinary AI (TransAI). IEEE, 2022, pp. 63–70.

Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nature machine intelligence* **2019**, *1*, 389–399.

Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; others. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* **2020**, *58*, 82–115.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* **2021**, *54*, 1–35.

Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; others. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and machines* **2018**, *28*, 689–707.

Mittelstadt, B.D.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data & Society* **2016**, *3*, 2053951716679679.

Radclyffe, C.; Ribeiro, M.; Wortham, R.H. The assessment list for trustworthy artificial intelligence: A review and recommendations. *Frontiers in artificial intelligence* **2023**, *6*, 1020592.

Lu, Q.; Zhu, L.; Xu, X.; Whittle, J.; Zowghi, D.; Jacquet, A. Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering. *ACM Computing Surveys* **2024**, *56*, 1–35.

Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine bias. In *Ethics of data and analytics*; Auerbach Publications, 2022; pp. 254–264.

Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics*; Auerbach Publications, 2022; pp. 296–299.

Wörsdörfer, M. The EU's artificial intelligence act: an ordoliberal assessment. *AI and Ethics* **2023**, pp. 1–16.

Khan, A.A.; Badshah, S.; Liang, P.; Waseem, M.; Khan, B.; Ahmad, A.; Fahmideh, M.; Niazi, M.; Akbar, M.A. Ethics of AI: A systematic literature review of principles and challenges. Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering, 2022, pp. 383–392.

Alzubaidi, L.; Al-Sabaawi, A.; Bai, J.; Dukhan, A.; Alkenani, A.H.; Al-Asadi, A.; Alwzwazy, H.A.; Manoufali, M.; Fadhel, M.A.; Albahri, A.; others. Towards Risk-Free Trustworthy Artificial Intelligence: Significance and Requirements. *International Journal of Intelligent Systems* **2023**, *2023*, 4459198.

Prem, E. From ethical AI frameworks to tools: a review of approaches. *AI and Ethics* **2023**, *3*, 699–716.

Liu, H.; Wang, Y.; Fan, W.; Liu, X.; Li, Y.; Jain, S.; Liu, Y.; Jain, A.; Tang, J. Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology* **2022**, *14*, 1–59.

Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* **2018**.

Blasch, E.; Pham, T.; Chong, C.Y.; Koch, W.; Leung, H.; Braines, D.; Abdelzaher, T. Machine learning/artificial intelligence for sensor data fusion–opportunities and challenges. *IEEE Aerospace and Electronic Systems Magazine* **2021**, *36*, 80–93.

Hendrycks, D.; Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136* **2016**.

Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* **2014**.

Raji, I.D.; Dobbe, R. Concrete problems in AI safety, revisited. *arXiv preprint arXiv:2401.10899* **2023**.

Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* **2017**.

Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.; Kagal, L. Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE, 2018, pp. 80–89.

Sweeney, L. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems* **2002**, *10*, 557–570.

Shokri, R.; Shmatikov, V. Privacy-preserving deep learning. Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, 2015, pp. 1310–1321.

Whittlestone, J.; Clark, J. Why and how governments should monitor AI development. *arXiv preprint arXiv:2108.12427* **2021**.

Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A.; Srikumar, M. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* **2020**.

Lundberg, S.; Lee, S.I. An unexpected unity among methods for interpreting model predictions. *arXiv preprint arXiv:1611.07478* **2016**.

Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. The limitations of deep learning in adversarial settings. 2016 IEEE European symposium on security and privacy (EuroS&P). IEEE, 2016, pp. 372–387.

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, 2016, pp. 308–318.

 doi:10.20944/preprints202410.0985.v1



Wei, K.; Li, J.; Ding, M.; Ma, C.; Su, H.; Zhang, B.; Poor, H.V. User-level privacy-preserving federated learning: Analysis and performance optimization. *IEEE Transactions on Mobile Computing* **2021**, *21*, 3388–3401.

Wagner, B.; d'Avila Garcez, A. Neural-symbolic integration for fairness in AI. CEUR Workshop Proceedings, 2021, Vol. 2846.

Amado, L.; Pereira, R.F.; Meneguzzi, F. Robust neuro-symbolic goal and plan recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, Vol. 37, pp. 11937–11944.

Pisano, G.; Ciatto, G.; Calegari, R.; Omicini, A.; others. Neuro-symbolic computation for XAI: Towards a unified model. CEUR WORKSHOP PROCEEDINGS. Sun SITE Central Europe, RWTH Aachen University, 2020, Vol. 2706, pp. 101–117.

Oltramari, A.; Francis, J.; Henson, C.; Ma, K.; Wickramarachchi, R. Neuro-symbolic architectures for context understanding. In *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*; IOS Press, 2020; pp. 143–160.

Venugopal, D.; Rus, V.; Shakya, A. Neuro-symbolic models: A scalable, explainable framework for strategy discovery from big edu-data. Proceedings of the 2nd Learner Data Institute Workshop in Conjunction with The 14th International Educational Data Mining Conference, 2021.

Himmelhuber, A.; Grimm, S.; Zillner, S.; Joblin, M.; Ringsquandl, M.; Runkler, T. Combining sub-symbolic and symbolic methods for explainability. Rules and Reasoning: 5th International Joint Conference, RuleML+ RR 2021, Leuven, Belgium, September 13–15, 2021, Proceedings 5. Springer, 2021, pp. 172–187.

Bellucci, M. Symbolic approaches for explainable artificial intelligence. PhD thesis, Normandie Université, 2023.

Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; others. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys* **2023**, *55*, 1–33.

Mileo, A. Towards a neuro-symbolic cycle for human-centered explainability. *Neurosymbolic Artificial Intelligence*, pp. 1–13.

Thota, S.R.; Arora, S. Neurosymbolic AI for Explainable Recommendations in Frontend UI Design-Bridging the Gap between Data-Driven and Rule-Based Approaches **2024**.

Xie, X.; Kersting, K.; Neider, D. Neuro-symbolic verification of deep neural networks. *arXiv preprint arXiv:2203.00938* **2022**.

Padalkar, P.; Ślusarz, N.; Komendantskaya, E.; Gupta, G. A Neurosymbolic Framework for Bias Correction in CNNs. *arXiv preprint arXiv:2405.15886* **2024**.

Hooshyar, D.; Yang, Y. Neural-symbolic computing: a step toward interpretable AI in education. *Bulletin of the Technical Committee on Learning Technology (ISSN: 2306-0212)* **2021**, *21*, 2–6.

Bennetot, A.; Franchi, G.; Del Ser, J.; Chatila, R.; Diaz-Rodriguez, N. Greybox XAI: A Neural-Symbolic learning framework to produce interpretable predictions for image classification. *Knowledge-Based Systems* **2022**, *258*, 109947.

Smirnova, A.; Yang, J.; Yang, D.; Cudre-Mauroux, P. Nessy: A Neuro-Symbolic System for Label Noise Reduction. *IEEE Transactions on Knowledge and Data Engineering* **2022**, *35*, 8300–8311.

Piplai, A.; Kotal, A.; Mohseni, S.; Gaur, M.; Mittal, S.; Joshi, A. Knowledge-enhanced neurosymbolic artificial intelligence for cybersecurity and privacy. *IEEE Internet Computing* **2023**, *27*, 43–48.

Zeng, Z. Neurosymbolic Learning and Reasoning for Trustworthy AI. PhD thesis, UCLA, 2024.

Agiollo, A.; Omicini, A. Measuring Trustworthiness in Neuro-Symbolic Integration. 2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS). IEEE, 2023, pp. 1–10.

Kosasih, E.; Papadakis, E.; Baryannis, G.; Brintrup, A. Explainable Artificial Intelligence in Supply Chain Management: A Systematic Review of Neurosymbolic Approaches. *International Journal of Production Research* **2023**.

Gaur, M.; Sheth, A. Building trustworthy NeuroSymbolic AI Systems: Consistency, reliability, explainability, and safety. *AI Magazine* **2024**, *45*, 139–155.

Selbst, A.D.; Boyd, D.; Friedler, S.A.; Venkatasubramanian, S.; Vertesi, J. Fairness and abstraction in sociotechnical systems. Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 59–68.

Christoph, M. *Interpretable machine learning: A guide for making black box models explainable*; Leanpub, 2020.

 doi:10.20944/preprints202410.0985.v1



Müller, V.C. Ethics of artificial intelligence and robotics. In *The Stanford Encyclopedia of Philosophy*; Stanford University, 2020.

Di Maio, P. Neurosymbolic knowledge representation for explainable and trustworthy ai **2020**.