Neuro-Symbolic methods for Trustworthy AI: a systematic review

Cyprien Michel–Delétie [a,b,*] and Md Kamruzzaman Sarker [c]
[a] *Computer Science Department, ENS de Lyon, France*
E-mail: cyprien.michel–deletie@ens-lyon.fr
[b] *Computer Science Department, University of Hartford, CT, USA*
[c] *Computer Science Department, Bowie State University, MD, USA*
E-mail: ksarker@bowiestate.edu

Abstract.
Recent advances in Artificial Intelligence (AI) especially in deep learning have manifested an increasing concern in trustworthiness, and its subparts such as interpretability, safety, fairness, and privacy. Neuro-symbolic methods, which mix some elements of neural networks with some elements of symbolic reasoning, have shown great potential for some aspects of trustworthiness. In this paper, we provide an overview of the various ways Neuro-Symbolic methods have been used to increase the trustworthiness, in the latest literature of the leading conferences. In particular, we focus on the contributions and limitations of the recent articles that discuss the interpretability, safety, fairness, and privacy of using the NeSy systems. We also did a categorization of the existing contributions along several key dimensions.

Keywords: Neuro-Symbolic, Trust, Interpretability

1. Introduction

The field of Artificial Intelligence (AI) is in a continuous state of exploration, with its potential applications appearing to be endless. AI decision-making systems have demonstrated superior performance, frequently outperforming humans. However, this comes with a notable drawback: the decision processes of these systems lack transparency and are often incomprehensible to humans. This issue becomes increasingly critical as AI systems begin to handle sensitive data and make crucial decisions in various sectors, ranging from autonomous driving to criminal justice. As a result, the demand for trustworthiness in AI systems is escalating. Particularly, the subject of interpretability has seen a significant rise in interest in recent years. This increase is a direct consequence of recognizing that many top-tier AI systems are non-transparent and difficult to interpret, leading them to be labeled as "black boxes". A common trend observed is that the larger the AI model, the more challenging it is to decipher its internal workings. These complex models pose a problem, as it becomes increasingly difficult to identify errors or biases within the system. Shifting towards more interpretable systems would cultivate greater trust in their decisions, enhance social acceptance, and encourage stakeholder discussions about their implementation [1].

Neuro-Symbolic AI, which combines Machine Learning and Knowledge Discovery and Data Mining (KDD), endeavors to integrate neural networks with symbolic processing techniques. This field attracts interest from two distinct perspectives [2]. From a cognitive science angle, while human brains exhibit connectionist characteristics

*Corresponding author. E-mail: cyprien.michel–deletie@ens-lyon.fr.

similar to neural networks, they also have the ability to process complex symbolic structures. This capability is believed to play a crucial role in the superiority of human intelligence over other animals. Additionally, it appears that symbolic and neural approaches complement each other, each with its own strengths and weaknesses. For example, deep learning systems, trained on raw data, show robustness against outliers, a feature less prominent in symbolic systems. In contrast, symbolic systems can directly utilize expert knowledge and are generally more self-explanatory compared to their neural counterparts.

The self-explanatory nature of Neuro-Symbolic methods is especially relevant when considering the aspect of trustworthiness. In this paper, we present a systematic review of recent literature (from 2021 to 2022) on Neuro-Symbolic approaches with a focus on achieving high trustworthiness. These studies were categorized based on their primary focus areas: privacy, fairness, safety, or interpretability. Notably, a majority of these papers concentrated on interpretability, and thus, they were further categorized using a traditional taxonomy in three dimensions: global versus local methods, self-explainable versus post-hoc explainability methods, and model-agnostic versus model-specific methods. This categorization provides an overview of the current trends in this domain, highlighting the areas that have been thoroughly explored and pinpointing promising directions for future research.

1.1. History of Neuro-Symbolic AI

The genesis of Neuro-Symbolic (NeSy) research is deeply intertwined with the history of Artificial Intelligence (AI), with its roots arguably dating back to a seminal 1943 paper by McCulloch and Pitts [3]. This pioneering work utilized propositional logic to model neural connections, setting the foundation for what would evolve into NeSy. Historically, the field of AI has been bifurcated into two primary paradigms: symbolism and connectionism. Symbolism approached intelligence through the lens of logic and rules, whereas connectionism favored learning driven by probabilistic methods. From the mid-1950s to the late 1980s, symbolic models dominated the early AI landscape, as researchers predominantly pursued this approach for crafting problem-solving systems [4]. However, the field encountered unforeseen hurdles, leading to the infamous "AI winter" of the 1980s, marked by a significant wane in AI interest and funding [5]. Despite this setback, research in symbolic AI persisted, albeit overshadowed by the resurgence of connectionist AI in the early 2010s. This revival, fueled by the impressive capabilities of deep learning in areas such as image classification, brought newfound attention to the field. Nevertheless, alongside these advancements came increasing concerns over the limitations of connectionist systems, such as vulnerability to adversarial attacks, low interpretability, challenges in integrating expert knowledge, and inherent biases.

NeSy emerged as a beacon of hope to address these challenges. While its conceptual roots extend back several decades, it was not until the 1990s that NeSy began to crystallize as a distinct field of study, gaining more structured research attention in the early 2000s [6]. NeSy aims to synthesize the strengths of both symbolic and neural elements, striving to create systems that exhibit robust learning capabilities (able to improve from raw data) and strong reasoning prowess (capable of abstraction and combinatorial reasoning). While neural networks have demonstrated impressive performance, logic remains a cornerstone in modeling thought and behavior [7]. The integration of these paradigms holds the promise of retaining their respective strengths while mitigating their weaknesses. However, this integration is challenging due to their fundamentally different methodologies: statistical inductive learning and distributed representations in connectionism, contrasted with logical deductive reasoning and localist representations in symbolism [4].

NeSy has shown its utility in various ways, such as leveraging symbolic knowledge bases and metadata to enhance deep learning systems, providing greater explainability through background knowledge, and solving complex problems that benefit from symbolic reasoning structures [2]. Susskind et al. demonstrated that NeSy could surpass purely deep learning approaches in visual question-answering tasks, achieving faster convergence and requiring less training data [5]. Additionally, NeSy has found successful applications in diverse industrial contexts, including business process modeling, trust management in e-commerce, coordination in large-scale multi-agent systems, and multi-modal processing and applications [7].

1.2. Background on Trustworthiness

The concept of trustworthiness is paramount in any decision-making system. At its core, a system is deemed trustworthy if it can be relied upon for high-stakes decisions with minimal or no supervision. While this certainly

encompasses performance, as a high-performing system is a prerequisite for trustworthiness, in the realm of AI, trustworthiness encompasses several additional dimensions: *interpretability*, *fairness*, *robustness* (to dataset shifts and data poisoning), *privacy*, and *safety* [8–10].

Fairness focuses on ensuring AI models do not harbor biases that could lead to discrimination against certain groups [11]. This is especially pertinent in AI applications involving people classification, such as risk assessments in criminal recidivism or automatic resume screening, both of which are rapidly gaining traction [12]. Studies have uncovered biases in some deployed systems against racial minorities, even in the absence of explicit racial data inputs. Addressing these biases to ensure fairness towards all groups is a critical concern.

Privacy relates to safeguarding the private data used in training AI models [8]. There is a risk that interaction with, or analysis of, the deployed model could inadvertently expose sensitive training data, a situation that raises significant privacy concerns.

Robustness pertains to the system's ability to function correctly in scenarios that deviate from its training data distribution [11]. This is vital across AI applications, as it is often impossible to anticipate all potential scenarios a system may encounter. Particularly concerning is the susceptibility of deep neural networks to adversarial attacks, where subtle manipulations of input data can lead to incorrect interpretations by the AI, despite being obvious to humans. Since robustness is intertwined with performance, it's not always clear when research specifically focuses on robustness; hence, papers primarily addressing robustness were not included in our review.

Safety is a critical aspect of trustworthiness that focuses on preventing accidents and unintended harmful behaviors in machine learning systems [10]. These issues can arise due to errors in specifying objectives, oversights in the learning process, or other implementation mistakes. As AI systems are increasingly deployed in complex and autonomous environments, ensuring their safety becomes paramount. This involves creating scalable solutions to mitigate risks and avoid potential adverse impacts on society, making AI systems not only effective but also reliable and secure.

Interpretability is the most extensively addressed aspect of trustworthiness, experiencing exponential growth as a research domain [13, 14]. There is little consensus on the definition of interpretability, but it can be broadly defined as the extent to which a system's operations can be understood by users [1, 15]. This includes access to mechanisms or reasoning that underpin the system's predictions. Interpretability and explainability are often used interchangeably, so we adopted a simplified definition treating them as synonymous. Simpler systems are naturally more interpretable, which is why this wasn't a major topic in earlier AI systems that used simpler methods like decision trees. However, with the complexity of deep neural networks, interpretability has become a critical concern, both for societal acceptance and regulatory compliance, with both the US and EU mandating a right to explanation for consumers [13]. We also argue that interpretability is crucial for a better understanding of the systems, which will help to develop them further and to overcome their flaws.

The characterization and approach to interpretability in AI is a subject of ongoing debate. While many papers use explainability and interpretability interchangeably, some argue that explainability is a stronger concept than interpretability [13, 16, 17]. Our review is based on the terminology used by the authors of the papers, which may not always align with this distinction. Generally, interpretability is self-assessed by researchers, leading to calls for more rigorous taxonomies and evaluations [15, 18, 19]. It's also important to note that explainability isn't the "silver bullet" for AI trustworthiness. Studies have shown that while explainability can enhance AI collaboration with novices, it doesn't necessarily do so with experts [20]; a combination of AI and human decision-making can be quicker but less accurate when AI provides explanations [19]; and there's a risk that explanations, even if not particularly useful, can unduly increase public acceptance, leading to overreliance on AI [21, 22].

Interpretability in AI systems has been tackled through a variety of methods. Some systems are inherently designed to be easily interpretable from the inside, termed as self-explainable or ante-hoc explainable methods. These systems are structured so that their internal processes are straightforward and clear. Another common approach is to create an interpretable layer for systems that are not inherently transparent, known as post-hoc explainability. This method is particularly versatile as it can be applied to virtually any system, allowing for the continued use of high-performance models. However, a drawback of post-hoc explainability is that the explanations it provides might not always accurately reflect the true workings of the system. This concern is highlighted by Rudin [23], who argues against the use of such explainability, suggesting that it can be misleading. Conversely, Gilpin et al. [17] propose that while using post-hoc explanations, it's crucial to clearly inform users about their potential limitations. There

are also approaches that fall somewhere between these two extremes. These methods aim to train systems in a way that makes their decision-making processes easier to interpret, without fundamentally altering their core structure.

In terms of the scope of explanations, they can range from local to global. Local explanations are tailored to individual instances, providing insight into specific decisions or similar cases. A well-known example of a local explanation method is LIME, which is designed to offer explanations for particular data points. On the other end of the spectrum, global explanations aim to shed light on the system's behavior as a whole, irrespective of individual inputs. Some methods provide explanations for a specific category of inputs, offering a more targeted understanding of the system's decisions in particular scenarios. These diverse approaches to interpretability demonstrate the complexity and varied nature of making AI systems transparent and understandable.

1.2.1. Link between Interpretability and NeSy

Neuro-symbolic AI (NeSy) and interpretability are intrinsically connected, primarily because symbols serve as an effective medium for explanations. Common practices in generating explanations include the use of decision trees or logic rules, which are inherently symbolic. Kambhampati et al. [24] have even suggested that symbols are essential for effective communication between humans and AI systems. While visual representations like saliency maps are also popular for explanations, these may not be adequate for complex human-AI interactions that require a blend of tacit and explicit task knowledge. Since NeSy inherently involves dealing with symbols within decision systems, it naturally possesses a strong potential for high interpretability.

Another perspective on the connection between NeSy and interpretability is their shared role as intermediaries linking deep learning with neuroscience. As Angelov et al. [13] have pointed out, a key objective of explainability is to mimic human-like reasoning in a manner that elucidates the predictions made by AI systems. This goal aligns closely with the principles of NeSy, which integrates aspects of human cognitive processes and neural network-based learning. Therefore, the synergy between NeSy and interpretability is not only practical in terms of implementing symbolic representations for explanations but also fundamental in achieving a deeper, more human-like understanding of AI decision-making processes.

2. Related Works

Trustworthiness, being a broad and multifaceted concept in AI, encompasses a diverse range of studies and reviews, often focused on specific domains within the field. A notable comprehensive survey by Liu, Wang et al. [25] addresses recent techniques for enhancing AI trustworthiness. This work examines trustworthiness across six dimensions: explainability, robustness, accountability & auditability, privacy, fairness, and environmental well-being. Another notable review in the realm of trustworthy Machine Learning was conducted by Serban et al. [26], providing valuable insights into methods for fostering trust in AI systems.

Interpretability methods in AI have received considerable attention, with numerous reviews dedicated to this topic. For instance, Speith et al. [14] conducted an analysis of various taxonomies used for categorizing interpretability methods. Their study revealed that these taxonomies are based on different criteria, such as the methods used, the type of explainability produced, or the conceptual approach, sometimes combining several of these aspects. Speith et al. argued that the choice of taxonomy should align with the user's needs and proposed a unified taxonomy to guide users. Reviews like that of Vilone et al. [27] have performed extensive classifications of explainable artificial methods, focusing on the formats of their outputs. This approach is highly beneficial for users seeking the most suitable system for their specific requirements. In contrast, our work is more geared towards researchers, considering that end-users may not be concerned with whether the model they use is neuro-symbolic.

Reviews specifically focusing on neuro-symbolic learning have also been published, including works by Sarker et al. [28], Besold et al. [7], Berlot-Attwell [29], and Hamilton et al. [30]. Sarker et al. provided a systematic review of Neuro-Symbolic methods presented in leading conference proceedings, applying two different taxonomies to categorize these methods and noting a recent increase in their popularity. Besold et al. presented a more subjective review of the neural-symbolic field, discussing its foundations, current applications, and future challenges. Berlot-Attwell explored the use of NeSy AI in Visual Question Answering (VQA), while Hamilton et al. offered a detailed analysis of NeSy methods in Natural Language Processing (NLP), highlighting the challenges in classifying papers as NeSy due to the term's ambiguity.

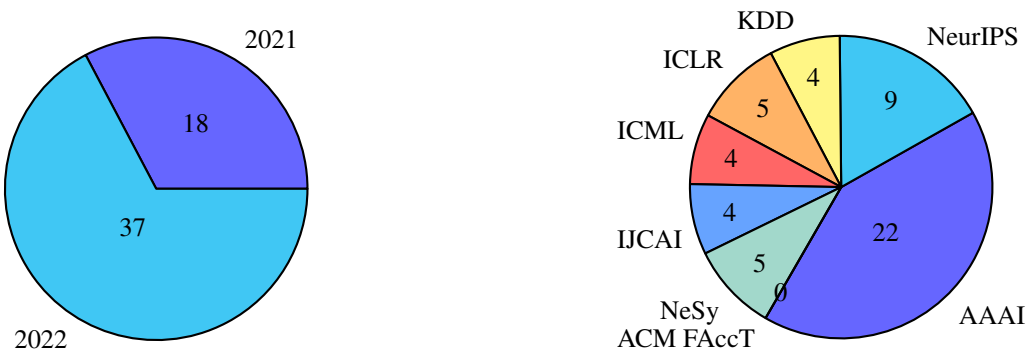
To our knowledge, this paper may be the first to review Neuro-Symbolic methods specifically through the lens of Trustworthy AI, marking a unique contribution to the field.

3. Survey methodology

This survey aims to capture the current state of research in the application of Neuro-Symbolic tools for enhancing trustworthiness in AI. We focused on papers published in top academic venues from 2021 to 2023, including those available up to May 2023. We selected papers from the following conferences: *NeurIPS*, *AAAI*, *IJCAI*, *IJCL*, *ICML*, *NeSy*, *AACM FAccT*, and *KDD*. The sheer volume of papers presented at these conferences, exceeding 10,000 over the last two years, necessitated a more strategic approach to identify relevant papers, rather than reviewing each one individually.

In our commitment to transparency, we employed a detailed and systematic methodology. Utilizing the *dblp* database, we initially filtered papers based on titles that contained keywords indicative of Neuro-Symbolic methods. These keywords included *symbol*, *logic* (excluding derivatives like *biologic* or *topologic*), *reason*, *inducti(on)*, *abducti(on)*, *concept*, *hybrid*, *ontolog(y)*, *relational*, *compositional*, and *rule*. We then used search-in-page tools to determine if these papers frequently mentioned key terms related to trustworthiness, such as *interpretab(le)*, *explaina(ble)*, *explanat(ion)*, *trust*, *fair*, *faithful*, *priva(cy)*, *tractab(le)*, and *understandab(le)*. Papers meeting these criteria were examined in more detail to assess their relevance to our focus.

Additionally, to ensure we did not overlook papers that explicitly mentioned the use of Neuro-Symbolic methods (but not in the title), we screened all papers with titles suggesting a focus on trustworthiness. We then reviewed papers that contained multiple mentions of the keyword *symbol* for further evaluation. This comprehensive approach was designed to capture a wide range of relevant research, ensuring a thorough overview of the intersection of Neuro-Symbolic methods and trustworthiness in AI.



(a) Distribution of papers by year

(b) Conference from which the papers where taken

Fig. 1. Distribution of selected papers

Determining whether a paper's approach qualifies as Neuro-Symbolic (NeSy) presented a significant challenge due to the broadness and ambiguity surrounding the definition of NeSy. To address this, we established specific criteria: a paper was included in our review only if it involved some form of symbolic knowledge manipulation (such as logic propositions, rules, action models, or graphs) directly contributing to trustworthiness. We specifically looked for papers where this symbolic knowledge played an active role in the process, rather than being a mere output. For example, if the explanations were presented in the form of a tree that was neither used nor executed in the system, we did not consider the method to be sufficiently neuro-symbolic.

While we recognize that this approach might have excluded some relevant papers, our objective was to minimize any systematic bias in our selection process that could lead to a skewed representation of the field. We noticed

that many papers treated interpretability as a beneficial byproduct rather than a primary focus, without substantial discussion or emphasis. To maintain the relevance and specificity of our survey, we chose to include only those papers where trustworthiness was a central motivation of the research. This decision inevitably introduced a degree of subjectivity into the selection of papers, but it was a necessary step to ensure the focus and coherence of our survey.

4. NeSy Research for Trustworthiness

Our comprehensive review yielded a total of 54 papers that employed neuro-symbolic (NeSy) methods with a clear emphasis on trustworthiness. An interesting pattern emerged from our analysis: the vast majority of these papers, except for two (one focusing on fairness [31] and another on safety [32]), concentrated on interpretability. This trend was notable despite our efforts to encompass a broader range of trustworthiness aspects such as fairness and privacy. This observation suggests that, currently, NeSy may not be widely utilized for addressing trustworthiness concerns beyond interpretability. Additionally, a significant increase in relevant publications was noted in 2022, with 37 out of the 54 papers coming from this year alone (Figure 1a), indicating a growing interest and expansion in this domain. The distribution of these papers across various conferences, as depicted in Figure 1b, reveals that AAAI is the predominant venue for this type of research.

4.1. Classification Based on Symbolic Data-Structures

In our categorization of the papers, we found that 16 of them focused on rule-learning approaches [33–48]. These papers typically utilize deep learning to generate logic rules or decision trees for classification purposes. In these instances, the symbolic component manifests as the output model, offering a high degree of transparency and interpretability. Beyond rule-learning approaches, we also analyzed the types of symbolic data structures employed in other NeSy systems. Excluding the rule-learning papers, we identified that these systems could be broadly categorized into three types based on the symbolic data structures they manipulate: logic structures, graphs, and other structures. This classification, depicted in Figure 2, provides insights into the varied approaches within the NeSy field, highlighting the diversity of methods being explored to enhance trustworthiness in AI systems.

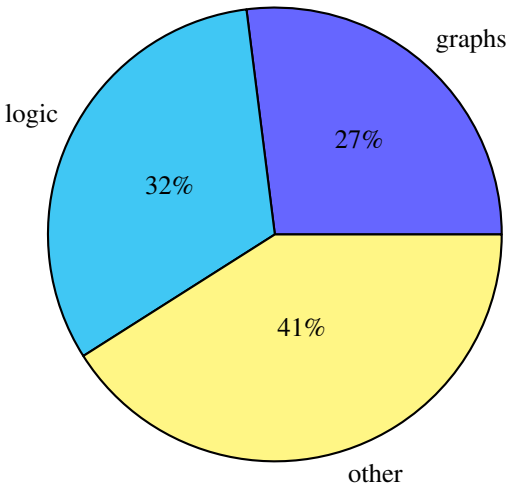


Fig. 2. Form of the symbolic knowledge (excluding rule learning papers)

Logic structures, as represented in various papers [31, 49–59], typically involve logic propositions, often in the form of logic rules (e.g., *precondition* $\rightarrow$ *class*). This approach uses symbolic reasoning as a means to interpret and classify data.

Graphs are another prevalent structure in NeSy research, encompassing a variety of types. Knowledge graphs are commonly used [60–63], but the category also includes other kinds of graphs [64–69], such as scene graphs, proof graphs, or Abstract Meaning Representation (AMR) graphs. These graphical structures are instrumental in representing relationships and dependencies in a visual and often intuitive manner.

The third category, labeled as "other," encompasses a variety of other symbolic data forms [32, 70–83]. This includes, for example, symbolic descriptions of objects or symbolic programming languages. This category is diverse and encompasses a wide range of approaches where symbolic representations take on various forms.

Interestingly, each of these three categories—logic structures, graphs, and other structures—encompasses a similar number of papers, illustrating the breadth and diversity of approaches within the neuro-symbolic field for enhancing trustworthiness. These varied methodologies highlight the versatility of symbolic representations in AI and their potential to address different aspects of trustworthiness in sophisticated and nuanced ways.

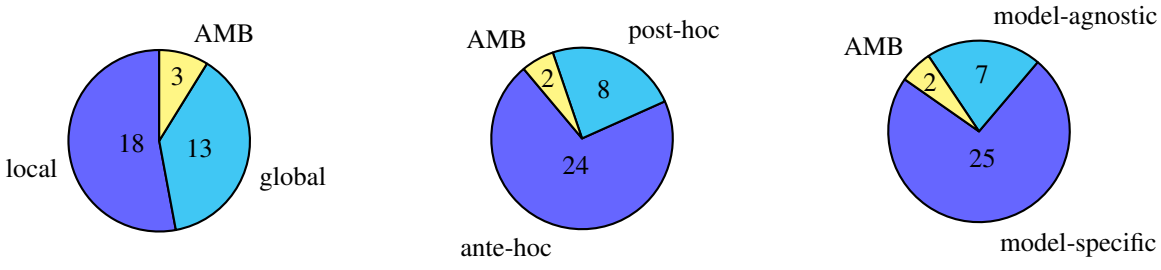
4.2. *Classification based on the types of interpretability*

To delve deeper into the papers that primarily focus on interpretability, which constitutes the bulk of our collection, we classified them based on three widely recognized dimensions in interpretability research: the scope of explainability, the stage at which the method is applied, and the method's dependence on the model (refer to table 1 for detailed classification). These dimensions are frequently used in analyzing papers in this field. In our analysis, we excluded rule-learning methods as they inherently fall into the ante-hoc, model-specific, and usually global scope categories.

The first dimension, the scope of explainability, differentiates between *local* and *global* explanations (Figure 3a). Local explanations are specific to a given input, providing insights into why a particular decision was made. On the other hand, global explanations offer a broader understanding, characterizing the behavior of the entire model. There's also an intermediate scope, which we might term as "cohort scope", applicable to a subset of inputs rather than just one or the entire model. Our review found a relatively balanced number of papers across these different scopes of explainability.

Regarding the stage of explanation, methods can be categorized as either *ante-hoc* (also known as *self-explainable*) or *post-hoc* (Figure 3b). Ante-hoc or self-explainable methods are designed to be inherently explainable, while post-hoc methods generate explanations after the fact, often for decisions made by an opaque, black-box model. Post-hoc explanations can take various forms, such as a textual justification of a decision or a simplified model that mirrors the original model's decisions.

The third dimension concerns whether the interpretability method is *model-agnostic* or *model-specific* (Figure 3c). Model-agnostic methods can be applied universally across different models, while model-specific methods are tailored to a particular model. Generally, post-hoc explainability methods have the flexibility to be model-agnostic. An interesting exception we noted is the work by Seungeon Lee [59], which involved modifying the final layer and training process of a deep model to enhance explainability. Our survey found no clear correlation between the scope of explanations and the stage at which the method is applied, indicating a diverse range of approaches in addressing interpretability in AI systems.



(a) Scope (b) Stage (c) Type

Fig. 3. Distribution of the different methods of explainability (AMB is for paper whose classification along this dimension is ambiguous)

Dimensions	(a)	(b)	AMB
local (a) vs global (b)	[50, 51, 53, 54, 61–67, 70, 73–75, 80, 82, 83]	[37, 51, 52, 55–58, 60, 68, 72, 76–78]	[49, 79, 81]
ante-hoc (a) vs post-hoc (b)	[37, 45, 50–53, 55, 58–63, 65–68, 70, 71, 73, 74, 77–79, 83]	[49, 54, 56, 72, 75, 80–82]	[57, 76]
model-specific (a) vs model-agnostic (b)	[37, 45, 49–53, 55, 58, 60–63, 65–68, 70, 71, 73, 74, 77–79, 81, 83],	[54, 56, 59, 72, 75, 80, 82]	[76, 81]

Table 1
Classification of the papers about explainability (AMB is for paper whose classification along this dimension is ambiguous)

4.3. Applications of NeSy Systems

Though we explored the trustworthiness contribution of the NeSy systems, while doing so, we found the NeSy systems were being used in different types of applications. Many of the systems proposed were working with visual data: either image classification [49, 54, 57, 70], action recognition in videos [74], agent communication about images [79], hand-written mathematical expression recognition [65], visual relation detection [71], or visual reasoning [78]. Equally many of the systems dealt with natural language settings: fake news detection [51, 66, 69], question answering [67, 68], unspecified NLP [81], text classification [31], commonsense reasoning [50], medical diagnosis through dialogue [77], text fiction tasks [63], or news recommendation [62]. A few applications were entirely based on graphs: knowledge graph completion [55, 60], query answering on knowledge graphs [61], graph classification [56], or imitating algorithms on graphs [52]. Some researchers worked in settings where an agent has to make different decisions (often reinforcement learning) [72, 73, 75, 80]. In some cases, the methods were explicitly suited for multiple settings [53, 65]. Lastly, a lot of other unique settings were explored: adaptive management [64], time series analysis [82], congestion control [58], safety execution of programs [32], or computer algebra [83]. The wide range of applications shows how versatile NeSy methods can be.

4.4. Lack of Research on Fairness, Privacy, and Safety

The primary aim of this review was to explore the application of Neuro-Symbolic (NeSy) systems in addressing various trustworthiness issues in AI. While we anticipated interpretability to be a predominant focus, the scarcity of research on NeSy systems related to fairness and privacy was notably surprising.

In the context of privacy, the potential benefits of incorporating NeSy systems are not immediately apparent. It could be suggested that NeSy may not offer significant advantages for enhancing privacy in AI systems. However, caution is advised before making definitive statements about NeSy's limitations in this area. As for fairness, there seems to be untapped potential for NeSy integration. For instance, the study by Wang et al. [84] approached fairness by imposing rule-like constraints during the training process. Although this approach was deemed too narrow to qualify as a comprehensive NeSy integration in our review, it indicates that integrating fairness constraints could be facilitated by NeSy models. This suggests that further investigation into NeSy's potential to address fairness in AI is warranted.

Another observation from our review is the wide range of methods encompassed under the NeSy umbrella and the absence of clear categorization for these methods. The term "neuro-symbolic" itself is often not explicitly used in many papers. While review papers like those by Sarker et al. [28] and Wang [4] propose conceptual taxonomies for NeSy systems, these classifications are not universally adopted in the literature. This lack of standardized taxonomy makes it challenging to categorize papers without a deep dive into their methodologies. Consequently, there is a need for more consensus in the research community regarding the taxonomy and terminology of NeSy systems. A more unified approach would facilitate the identification and comparison of works with similar methodologies, regardless of their specific applications.

5. Conclusion

This research endeavor embarked on a comprehensive examination of the most recent advancements in Neuro-Symbolic (NeSy) methods, specifically focusing on their role in enhancing the trustworthiness of AI systems. Our findings reveal that the primary application of NeSy methods in current research is centered around augmenting interpretability in AI. By converging the fields of AI trustworthiness and NeSy integration, this study pioneers a unified analysis of these two intertwined domains.

The papers included in our review were systematically categorized based on the scope, stage, and adaptability of the interpretability methods they employed. A key insight from our study is the recognition of the immense potential NeSy integration holds in the realm of interpretability, applicable across a myriad of settings. This potential is not constrained by any specific domain or application, indicating a broad and versatile utility of NeSy approaches.

However, our study also highlights a noteworthy imbalance in the focus of current NeSy research. While a substantial part of this research is dedicated to enhancing interpretability, there is a noticeably smaller portion of work aimed at improving other aspects of AI trustworthiness, such as security. This observation underscores an opportunity for future research to broaden the scope of NeSy applications, extending its benefits to other critical dimensions of AI trustworthiness, including but not limited to fairness, privacy, and safety.

In conclusion, our review establishes a foundational understanding of the current state of NeSy research in the context of AI trustworthiness, particularly interpretability, and opens avenues for future exploration in expanding the application of NeSy methods to address a wider array of trustworthiness concerns in AI systems.

References

[1] A. Kasirzadeh, Reasons, Values, Stakeholders: A Philosophical Framework for Explainable Artificial Intelligence, in: *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, M.C. Elish, W. Isaac and R.S. Zemel, eds, ACM, 2021, p. 14. doi:10.1145/3442188.3445866.

[2] P. Hitzler, A. Eberhart, M. Ebrahimi, M.K. Sarker and L. Zhou, Neuro-symbolic approaches in artificial intelligence, *National Science Review* **9**(6) (2022), nwac035. doi:10.1093/nsr/nwac035.

[3] W.S. McCulloch and W. Pitts, A logical calculus of the ideas immanent in nervous activity, *The Bulletin of Mathematical Biophysics* **5**(4) (1943), 115–133. doi:10.1007/bf02478259.

[4] W. Wang, Y. Yang and F. Wu, Towards Data-and Knowledge-Driven Artificial Intelligence: A Survey on Neuro-Symbolic Computing, arXiv, 2022. doi:10.48550/ARXIV.2210.15889. https://arxiv.org/abs/2210.15889.

[5] Z. Susskind, B. Arden, L.K. John, P. Stockton and E.B. John, Neuro-Symbolic AI: An Emerging Class of AI Workloads and their Characterization, *CoRR* **abs/2109.06133** (2021). https://arxiv.org/abs/2109.06133.

[6] S. Shi, H. Chen, W. Ma, J. Mao, M. Zhang and Y. Zhang, Neural Logic Reasoning, in: *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, M. d'Aquin, S. Dietze, C. Hauff, E. Curry and P. Cudré-Mauroux, eds, ACM, 2020, pp. 1365–1374. doi:10.1145/3340531.3411949.

[7] T.R. Besold, A.S. d'Avila Garcez, S. Bader, H. Bowman, P.M. Domingos, P. Hitzler, K. Kühnberger, L.C. Lamb, D. Lowd, P.M.V. Lima, L. de Penning, G. Pinkas, H. Poon and G. Zaverucha, Neural-Symbolic Learning and Reasoning: A Survey and Interpretation, *CoRR* **abs/1711.03902** (2017). http://arxiv.org/abs/1711.03902.

[8] N. Papernot, What does it mean for ML to be trustworthy?, ICML Workshop on Participatory Approaches to Machine Learning, 2020. https://youtu.be/UpGgIqLhaqo.

[9] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi and B. Zhou, Trustworthy AI: From Principles to Practices, *CoRR* **abs/2110.01167** (2021). https://arxiv.org/abs/2110.01167.

[10] D. Amodei, C. Olah, J. Steinhardt, P.F. Christiano, J. Schulman and D. Mané, Concrete Problems in AI Safety, *CoRR* **abs/1606.06565** (2016). http://arxiv.org/abs/1606.06565.

[11] K.R. Varshney, Trustworthy machine learning and artificial intelligence, *XRDS* **25**(3) (2019), 26–29. doi:10.1145/3313109.

[12] J. Schoeffer, N. Kuehl and Y. Machowski, "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making, in: *2022 ACM Conference on Fairness, Accountability, and Transparency*, ACM, 2022. doi:10.1145/3531146.3533218.

[13] P.P. Angelov, E.A. Soares, R. Jiang, N.I. Arnold and P.M. Atkinson, Explainable artificial intelligence: an analytical review, *WIREs Data Mining Knowl. Discov.* **11**(5) (2021). doi:10.1002/widm.1424.

[14] T. Speith, A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods, in: *2022 ACM Conference on Fairness, Accountability, and Transparency*, ACM, 2022. doi:10.1145/3531146.3534639.

[15] F. Doshi-Velez and B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, 2017.

[16] A. Bell, I. Solano-Kamaiko, O. Nov and J. Stoyanovich, It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy, in: *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, ACM, 2022, pp. 248–266. doi:10.1145/3531146.3533090.

[17] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M.A. Specter and L. Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, in: *5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018*, F. Bonchi, F.J. Provost, T. Eliassi-Rad, W. Wang, C. Cattuto and R. Ghani, eds, IEEE, 2018, pp. 80–89. doi:10.1109/DSAA.2018.00018.

[18] K. Sokol and P.A. Flach, Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches, *CoRR* **abs/1912.05100** (2019). http://arxiv.org/abs/1912.05100.

[19] S. Jesus, C. Belém, V. Balayan, J. Bento, P. Saleiro, P. Bizarro and J. Gama, How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations (2021). doi:10.48550/ARXIV.2101.08758. https://arxiv.org/abs/2101.08758.

[20] R.R. Paleja, M. Ghuy, N.R. Arachchige, R. Jensen and M.C. Gombolay, The Utility of Explainable AI in Ad Hoc Human-Machine Teaming, in: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y.N. Dauphin, P. Liang and J.W. Vaughan, eds, 2021, pp. 610–623. https://proceedings.neurips.cc/paper/2021/hash/05d74c48b5b30514d8e9bd60320fc8f6-Abstract.html.

[21] A. Weller, Transparency: Motivations and Challenges, in: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen and K. Müller, eds, Lecture Notes in Computer Science, Vol. 11700, Springer, 2019, pp. 23–40. doi:10.1007/978-3-030-28954-6_2.

[22] A. Ferrario and M. Loi, How Explainability Contributes to Trust in AI, in: *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, ACM, 2022, pp. 1457–1466. doi:10.1145/3531146.3533202.

[23] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nat. Mach. Intell.* **1**(5) (2019), 206–215. doi:10.1038/s42256-019-0048-x.

[24] S. Kambhampati, S. Sreedharan, M. Verma, Y. Zha and L. Guan, Symbols as a Lingua Franca for Bridging Human-AI Chasm for Explainable and Advisable AI Systems, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 12262–12267. https://ojs.aaai.org/index.php/AAAI/article/view/21488.

[25] H. Liu, Y. Wang, W. Fan, X. Liu, Y. Li, S. Jain, Y. Liu, A. Jain and J. Tang, Trustworthy AI: A Computational Perspective, *ACM Trans. Intell. Syst. Technol.* **14**(1) (2022). doi:10.1145/3546872.

[26] A. Serban, K. van der Blom, H.H. Hoos and J. Visser, Practices for Engineering Trustworthy Machine Learning Applications, in: *1st IEEE/ACM Workshop on AI Engineering - Software Engineering for AI, WAIN@ICSE 2021, Madrid, Spain, May 30-31, 2021*, IEEE, 2021, pp. 97–100. doi:10.1109/WAIN52551.2021.00021.

[27] G. Vilone and L. Longo, Classification of Explainable Artificial Intelligence Methods through Their Output Formats, *Mach. Learn. Knowl. Extr.* **3**(3) (2021), 615–661. doi:10.3390/make3030032.

[28] M.K. Sarker, L. Zhou, A. Eberhart and P. Hitzler, Neuro-Symbolic Artificial Intelligence: Current Trends, *CoRR* **abs/2105.05330** (2021). https://arxiv.org/abs/2105.05330.

[29] I. Berlot-Attwell, Neuro-Symbolic VQA: A review from the perspective of AGI desiderata, *CoRR* **abs/2104.06365** (2021). https://arxiv.org/abs/2104.06365.

[30] K. Hamilton, A. Nayak, B. Bozic and L. Longo, Is Neuro-Symbolic AI Meeting its Promise in Natural Language Processing? A Structured Review, *CoRR* **abs/2202.12205** (2022). https://arxiv.org/abs/2202.12205.

[31] H. Yao, Y. Chen, Q. Ye, X. Jin and X. Ren, Refining Language Models with Compositional Explanations, in: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y.N. Dauphin, P. Liang and J.W. Vaughan, eds, 2021, pp. 8954–8967. https://proceedings.neurips.cc/paper/2021/hash/4b26dc4663ccf960c8538d595d0a1d3a-Abstract.html.

[32] C. Yang and S. Chaudhuri, Safe Neurosymbolic Learning with Differentiable Symbolic Execution, in: *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022. https://openreview.net/forum?id=NYBmJN4MyZ.

[33] Z. Wang, W. Zhang, N. Liu and J. Wang, Scalable Rule-Based Representation Learning for Interpretable Classification, in: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y.N. Dauphin, P. Liang and J.W. Vaughan, eds, 2021, pp. 30479–30491. https://proceedings.neurips.cc/paper/2021/hash/ffbd6cbb019a1413183c8d08f2929307-Abstract.html.

[34] F. Yang, K. He, L. Yang, H. Du, J. Yang, B. Yang and L. Sun, Learning Interpretable Decision Rule Sets: A Submodular Optimization Approach, in: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y.N. Dauphin, P. Liang and J.W. Vaughan, eds, 2021, pp. 27890–27902. https://proceedings.neurips.cc/paper/2021/hash/eaa32c96f620053cf442ad32258076b9-Abstract.html.

[35] M. Landajuela, B.K. Petersen, S. Kim, C.P. Santiago, R. Glatt, T.N. Mundhenk, J.F. Pettit and D.M. Faissol, Discovering symbolic policies with deep reinforcement learning, in: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, M. Meila and T. Zhang, eds, Proceedings of Machine Learning Research, Vol. 139, PMLR, 2021, pp. 5979–5989. http://proceedings.mlr.press/v139/landajuela21a.html.

[36] M. Qu, J. Chen, L.A.C. Xhonneux, Y. Bengio and J. Tang, RNNLogic: Learning Logic Rules for Reasoning on Knowledge Graphs, in: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021. https://openreview.net/forum?id=tGZu6DlbreV.

[37] A. Kakadiya, S. Natarajan and B. Ravindran, Relational Boosted Bandits, in: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 12123–12130. https://ojs.aaai.org/index.php/AAAI/article/view/17439.

[38] M. Shvo, A.C. Li, R.T. Icarte and S.A. McIlraith, Interpretable Sequence Classification via Discrete Optimization, in: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 9647–9656. https://ojs.aaai.org/index.php/AAAI/article/view/17161.

[39] N. Topin, S. Milani, F. Fang and M. Veloso, Iterative Bounding MDPs: Learning Interpretable Policies via Non-Interpretable Methods, in: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 9923–9931. https://ojs.aaai.org/index.php/AAAI/article/view/17192.

[40] R.K. Yadav, L. Jiao, O. Granmo and M. Goodwin, Human-Level Interpretable Learning for Aspect-Based Sentiment Analysis, in: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 14203–14212. https://ojs.aaai.org/index.php/AAAI/article/view/17671.

[41] A. Dhaou, A. Bertoncello, S. Gourvénec, J. Garnier and E.L. Pennec, Causal and Interpretable Rules for Time Series Analysis, in: *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, F. Zhu, B.C. Ooi and C. Miao, eds, ACM, 2021, pp. 2764–2772. doi:10.1145/3447548.3467161.

[42] C. Glanois, Z. Jiang, X. Feng, P. Weng, M. Zimmer, D. Li, W. Liu and J. Hao, Neuro-Symbolic Hierarchical Rule Induction, in: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu and S. Sabato, eds, Proceedings of Machine Learning Research, Vol. 162, PMLR, 2022, pp. 7583–7615. https://proceedings.mlr.press/v162/glanois22a.html.

[43] S. Li, M. Feng, L. Wang, A. Essofi, Y. Cao, J. Yan and L. Song, Explaining Point Processes by Learning Interpretable Temporal Logic Rules, in: *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022. https://openreview.net/forum?id=P07dq7iSAGr.

[44] P. Sen, B.W.S.R. de Carvalho, R. Riegel and A.G. Gray, Neuro-Symbolic Inductive Logic Programming with Logical Neural Networks, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 8212–8219. https://ojs.aaai.org/index.php/AAAI/article/view/20795.

[45] Y. Yang, J.C. Kerce and F. Fekri, LOGICDEF: An Interpretable Defense Framework against Adversarial Examples via Inductive Scene Graph Reasoning, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 8840–8848. https://ojs.aaai.org/index.php/AAAI/article/view/20865.

[46] R.K. Yadav, L. Jiao, O. Granmo and M. Goodwin, Robust Interpretable Text Classification against Spurious Correlations Using AND-rules with Negation, in: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, L.D. Raedt, ed., ijcai.org, 2022, pp. 4439–4446. doi:10.24963/ijcai.2022/616.

[47] X. Liu, W. Lei, J. Lv and J. Zhou, Abstract Rule Learning for Paraphrase Generation, in: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, L.D. Raedt, ed., ijcai.org, 2022, pp. 4273–4279. doi:10.24963/ijcai.2022/593.

[48] M. Glauer, R. West, S. Michie and J. Hastings, ESC-Rules: Explainable, Semantically Constrained Rule Sets, in: *Proceedings of the 16th International Workshop on Neural-Symbolic Learning and Reasoning as part of the 2nd International Joint Conference on Learning & Reasoning (IJCLR 2022), Cumberland Lodge, Windsor Great Park, UK, September 28-30, 2022*, A.S. d'Avila Garcez and E. Jiménez-Ruiz, eds, CEUR Workshop Proceedings, Vol. 3212, CEUR-WS.org, 2022, pp. 94–103. https://ceur-ws.org/Vol-3212/paper7.pdf.

[49] M. de Sousa Ribeiro and J. Leite, Aligning Artificial Neural Networks and Ontologies towards Explainable AI, in: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 4932–4940. https://ojs.aaai.org/index.php/AAAI/article/view/16626.

[50] A. Kalyanpur, T. Breloff and D.A. Ferrucci, Braid: Weaving Symbolic and Neural Knowledge into Coherent Logical Explanations, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 10867–10874. https://ojs.aaai.org/index.php/AAAI/article/view/21333.

[51] J. Chen, Q. Bao, C. Sun, X. Zhang, J. Chen, H. Zhou, Y. Xiao and L. Li, LOREN: Logic-Regularized Reasoning for Interpretable Fact Verification, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 10482–10491. https://ojs.aaai.org/index.php/AAAI/article/view/21291.

[52] D. Georgiev, P. Barbiero, D. Kazhdan, P. Velickovic and P. Lió, Algorithmic Concept-Based Explainable Reasoning, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 6685–6693. https://ojs.aaai.org/index.php/AAAI/article/view/20623.

[53] P. Barbiero, G. Ciravegna, F. Giannini, P. Lió, M. Gori and S. Melacci, Entropy-Based Logic Explanations of Neural Networks, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 6046–6054. https://ojs.aaai.org/index.php/AAAI/article/view/20551.

[54] J. An, Y. Lai and Y. Han, Logic Rule Guided Attribution with Dynamic Ablation, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 77–85. https://ojs.aaai.org/index.php/AAAI/article/view/19881.

[55] D.J.T. Cucala, B.C. Grau, E.V. Kostylev and B. Motik, Explainable GNN-Based Models over Knowledge Graphs, in: *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022. https://openreview.net/forum?id=CrCvGNHAIrz.

[56] A. Himmelhuber, S. Zillner, S. Grimm, M. Ringsquandl, M. Joblin and T.A. Runkler, A New Concept for Explaining Graph Neural Networks, in: *Proceedings of the 15th International Workshop on Neural-Symbolic Learning and Reasoning as part of the 1st International Joint Conference on Learning & Reasoning (IJCLR 2021), Virtual conference, October 25-27, 2021*, A.S. d'Avila Garcez and E. Jiménez-Ruiz, eds, CEUR Workshop Proceedings, Vol. 2986, CEUR-WS.org, 2021, pp. 1–5. https://ceur-ws.org/Vol-2986/paper1.pdf.

[57] T. Kasioumis, J. Townsend and H. Inakoshi, Elite BackProp: Training Sparse Interpretable Neurons, in: *Proceedings of the 15th International Workshop on Neural-Symbolic Learning and Reasoning as part of the 1st International Joint Conference on Learning & Reasoning (IJCLR 2021), Virtual conference, October 25-27, 2021*, A.S. d'Avila Garcez and E. Jiménez-Ruiz, eds, CEUR Workshop Proceedings, Vol. 2986, CEUR-WS.org, 2021, pp. 82–93. https://ceur-ws.org/Vol-2986/paper6.pdf.

[58] S.P. Sharan, W. Zheng, K. Hsu, J. Xing, A. Chen and Z. Wang, Symbolic Distillation for Learned TCP Congestion Control, in: *NeurIPS*, 2022. http://papers.nips.cc/paper_files/paper/2022/hash/4574ac9854d4defe3bf119d07b817084-Abstract-Conference.html.

[59] S. Lee, X. Wang, S. Han, X. Yi, X. Xie and M. Cha, Self-explaining deep models with logic rule reasoning, in: *NeurIPS*, 2022. http://papers.nips.cc/paper_files/paper/2022/hash/1548d98b62d3a4382a31ba77d89186cd-Abstract-Conference.html.

[60] H. Zha, Z. Chen and X. Yan, Inductive Relation Prediction by BERT, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 5923–5931. doi:10.1609/AAAI.V36I5.20537. https://doi.org/10.1609/aaai.v36i5.20537.

[61] Z. Zhu, M. Galkin, Z. Zhang and J. Tang, Neural-Symbolic Models for Logical Queries on Knowledge Graphs, in: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu and S. Sabato, eds, Proceedings of Machine Learning Research, Vol. 162, PMLR, 2022, pp. 27454–27478. https://proceedings.mlr.press/v162/zhu22c.html.

[62] D. Liu, J. Lian, Z. Liu, X. Wang, G. Sun and X. Xie, Reinforced Anchor Knowledge Graph Generation for News Recommendation Reasoning, in: *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, F. Zhu, B.C. Ooi and C. Miao, eds, ACM, 2021, pp. 1055–1065. doi:10.1145/3447548.3467315.

[63] X. Peng, M.O. Riedl and P. Ammanabrolu, Inherently Explainable Reinforcement Learning in Natural Language, in: *NeurIPS*, 2022. http://papers.nips.cc/paper_files/paper/2022/hash/672e44a114a41d5f34b97459877c083d-Abstract-Conference.html.

[64] J. Ferrer-Mestres, T.G. Dietterich, O. Buffet and I. Chades, K-N-MOMDPs: Towards Interpretable Solutions for Adaptive Management, in: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 14775–14784. https://ojs.aaai.org/index.php/AAAI/article/view/17735.

[65] J. Wu, F. Yin, Y. Zhang, X. Zhang and C. Liu, Graph-to-Graph: Towards Accurate and Interpretable Online Handwritten Mathematical Expression Recognition, in: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 2925–2933. https://ojs.aaai.org/index.php/AAAI/article/view/16399.

[66] Y. Jin, X. Wang, R. Yang, Y. Sun, W. Wang, H. Liao and X. Xie, Towards Fine-Grained Reasoning for Fake News Detection, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 5746–5754. https://ojs.aaai.org/index.php/AAAI/article/view/20517.

[67] W. Zhong, J. Huang, Q. Liu, M. Zhou, J. Wang, J. Yin and N. Duan, Reasoning over Hybrid Chain for Table-and-Text Open Domain Question Answering, in: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, L.D. Raedt, ed., ijcai.org, 2022, pp. 4531–4537. doi:10.24963/ijcai.2022/629.

[68] Z. Deng, Y. Zhu, Y. Chen, M. Witbrock and P. Riddle, Interpretable AMR-Based Question Decomposition for Multi-hop Question Answering, in: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, L.D. Raedt, ed., ijcai.org, 2022, pp. 4093–4099. doi:10.24963/ijcai.2022/568.

[69] R. Yang, X. Wang, Y. Jin, C. Li, J. Lian and X. Xie, Reinforcement Subgraph Reasoning for Fake News Detection, in: *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, A. Zhang and H. Rangwala, eds, ACM, 2022, pp. 2253–2262. doi:10.1145/3534678.3539277.

[70] S. Jang, M.J.A. Girard and A.H. Thiéry, Explainable Diabetic Retinopathy Classification Based on Neural-Symbolic Learning, in: *Proceedings of the 15th International Workshop on Neural-Symbolic Learning and Reasoning as part of the 1st International Joint Conference on Learning & Reasoning (IJCLR 2021), Virtual conference, October 25-27, 2021*, A.S. d'Avila Garcez and E. Jiménez-Ruiz, eds, CEUR Workshop Proceedings, Vol. 2986, CEUR-WS.org, 2021, pp. 104–114. https://ceur-ws.org/Vol-2986/paper8.pdf.

[71] K. Chen and K.D. Forbus, Visual Relation Detection using Hybrid Analogical Learning, in: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 801–808. https://ojs.aaai.org/index.php/AAAI/article/view/16162.

[72] P. Verma, S.R. Marpally and S. Srivastava, Asking the Right Questions: Learning Interpretable Action Models Through Query Answering, in: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 12024–12033. https://ojs.aaai.org/index.php/AAAI/article/view/17428.

[73] M. Jin, Z. Ma, K. Jin, H.H. Zhuo, C. Chen and C. Yu, Creativity of AI: Automatic Symbolic Option Discovery for Facilitating Deep Reinforcement Learning, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 7042–7050. https://ojs.aaai.org/index.php/AAAI/article/view/20663.

[74] H. Hua, D. Li, R. Li, P. Zhang, J. Renz and A.G. Cohn, Towards Explainable Action Recognition by Salient Qualitative Spatial Object Relation Chains, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 5710–5718. https://ojs.aaai.org/index.php/AAAI/article/view/20513.

[75] S. Sreedharan, U. Soni, M. Verma, S. Srivastava and S. Kambhampati, Bridging the Gap: Providing Post-Hoc Symbolic Explanations for Sequential Decision-Making Problems with Inscrutable Representations, in: *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022. https://openreview.net/forum?id=o-1v9hdSult.

[76] A. Geiger, Z. Wu, H. Lu, J. Rozner, E. Kreiss, T. Icard, N.D. Goodman and C. Potts, Inducing Causal Structure for Interpretable Neural Networks, in: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu and S. Sabato, eds, Proceedings of Machine Learning Research, Vol. 162, PMLR, 2022, pp. 7324–7338. https://proceedings.mlr.press/v162/geiger22a.html.

[77] W. Liu, Y. Cheng, H. Wang, J. Tang, Y. Liu, R. Zhao, W. Li, Y. Zheng and X. Liang, "My nose is running." "Are you also coughing?": Building A Medical Diagnosis Agent with Interpretable Inquiry Logics, in: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, L.D. Raedt, ed., ijcai.org, 2022, pp. 4266–4272. doi:10.24963/ijcai.2022/592.

[78] M. Ding, Z. Chen, T. Du, P. Luo, J. Tenenbaum and C. Gan, Dynamic Visual Reasoning by Learning Differentiable Physics Models from Video and Language, in: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y.N. Dauphin, P. Liang and J.W. Vaughan, eds, 2021, pp. 887–899. https://proceedings.neurips.cc/paper/2021/hash/07845cd9aefa6cde3f8926d25138a3a2-Abstract.html.

[79] R. Dessì, E. Kharitonov and M. Baroni, Interpretable agent communication from scratch (with a generic visual processor emerging on the side), in: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y.N. Dauphin, P. Liang and J.W. Vaughan, eds, 2021, pp. 26937–26949. https://proceedings.neurips.cc/paper/2021/hash/e250c59336b505ed411d455abaa30b4d-Abstract.html.

[80] M. Finkelstein, N.L. Schlot, L. Liu, Y. Kolumbus, D.C. Parkes, J.S. Rosenschein and S. Keren, Explainable Reinforcement Learning via Model Transforms, in: *NeurIPS*, 2022. http://papers.nips.cc/paper_files/paper/2022/hash/dbef234be68d8b170240511639610fd1-Abstract-Conference.html.

[81] D. Zhang, H. Zhang, H. Zhou, X. Bao, D. Huo, R. Chen, X. Cheng, M. Wu and Q. Zhang, Building Interpretable Interaction Trees for Deep NLP Models, in: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 14328–14337. https://ojs.aaai.org/index.php/AAAI/article/view/17685.

[82] D. Rajapaksha and C. Bergmeir, LIMREF: Local Interpretable Model Agnostic Rule-Based Explanations for Forecasting, with an Application to Electricity Smart Meter Data, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 12098–12107. https://ojs.aaai.org/index.php/AAAI/article/view/21469.

[83] S. Peng, D. Fu, Y. Cao, Y. Liang, G. Xu, L. Gao and Z. Tang, Compute Like Humans: Interpretable Step-by-step Symbolic Computation with Deep Neural Network, in: *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, A. Zhang and H. Rangwala, eds, ACM, 2022, pp. 1348–1357. doi:10.1145/3534678.3539276.

[84] N. Wang, S. Nie, Q. Wang, Y. Wang, M. Sanjabi, J. Liu, H. Firooz and H. Wang, COFFEE: Counterfactual Fairness for Personalized Text Generation in Explainable Recommendation, *CoRR* **abs/2210.15500** (2022). doi:10.48550/arXiv.2210.15500.

[85] Z.C. Lipton, The mythos of model interpretability, *Commun. ACM* **61**(10) (2018), 36–43. doi:10.1145/3233231.

[86] A. Ignatiev, J. Marques-Silva, N. Narodytska and P.J. Stuckey, Reasoning-Based Learning of Interpretable ML Models, in: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, Z. Zhou, ed., ijcai.org, 2021, pp. 4458–4465. doi:10.24963/ijcai.2021/608.

[87] L.D. Raedt, R. Manhaeve, S. Dumancic, T. Demeester and A. Kimmig, Neuro-Symbolic = Neural + Logical + Probabilistic, in: *Proceedings of the 2019 International Workshop on Neural-Symbolic Learning and Reasoning (NeSy 2019), Annual workshop of the Neural-Symbolic Learning and Reasoning Association, Macao, China, August 12, 2019*, D. Doran, A.S. d'Avila Garcez and F. Lécué, eds, 2019.

[88] L.C. Lamb, A.S. d'Avila Garcez, M. Gori, M.O.R. Prates, P.H.C. Avelar and M.Y. Vardi, Graph Neural Networks Meet Neural-Symbolic Computing: A Survey and Perspective, in: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, C. Bessiere, ed., ijcai.org, 2020, pp. 4877–4884. doi:10.24963/ijcai.2020/679.

[89] V. Belle, Symbolic Logic meets Machine Learning: A Brief Survey in Infinite Domains, *CoRR* **abs/2006.08480** (2020). https://arxiv.org/abs/2006.08480.

[90] K. Chen and K.D. Forbus, Visual Relation Detection using Hybrid Analogical Learning, in: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 801–808. https://ojs.aaai.org/index.php/AAAI/article/view/16162.

[91] G.J. Stein, Generating High-Quality Explanations for Navigation in Partially-Revealed Environments, in: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y.N. Dauphin, P. Liang and J.W. Vaughan, eds, 2021, pp. 17493–17506. https://proceedings.neurips.cc/paper/2021/hash/926ec030f29f83ce5318754fdb631a33-Abstract.html.

[92] R. Kusters, Y. Kim, M. Collery, C. de Sainte Marie and S. Gupta, Differentiable Rule Induction with Learned Relational Features, in: *Proceedings of the 16th International Workshop on Neural-Symbolic Learning and Reasoning as part of the 2nd International Joint Conference on Learning & Reasoning (IJCLR 2022), Cumberland Lodge, Windsor Great Park, UK, September 28-30, 2022*, A.S. d'Avila Garcez and E. Jiménez-Ruiz, eds, CEUR Workshop Proceedings, Vol. 3212, CEUR-WS.org, 2022, pp. 30–44. https://ceur-ws.org/Vol-3212/paper3.pdf.

[93] J. Huang and K.C. Chang, Towards Reasoning in Large Language Models: A Survey, *CoRR* **abs/2212.10403** (2022). doi:10.48550/arXiv.2212.10403.

[94] N. Heist and H. Paulheim, The CaLiGraph Ontology as a Challenge for OWL Reasoners, in: *Proceedings of the Semantic Reasoning Evaluation Challenge (SemREC 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Event, October 27th, 2021*, G. Singh, R. Mutharaju and P. Kapanipathi, eds, CEUR Workshop Proceedings, Vol. 3123, CEUR-WS.org, 2021, pp. 21–31. https://ceur-ws.org/Vol-3123/paper3.pdf.

[95] S. Hao, Y. Gu, H. Ma, J.J. Hong, Z. Wang, D.Z. Wang and Z. Hu, Reasoning with Language Model is Planning with World Model, *CoRR* **abs/2305.14992** (2023). doi:10.48550/arXiv.2305.14992.

[96] S.M. Kazemi, N. Kim, D. Bhatia, X. Xu and D. Ramachandran, LAMBADA: Backward Chaining for Automated Reasoning in Natural Language, *CoRR* **abs/2212.13894** (2022). doi:10.48550/arXiv.2212.13894.

[97] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, E. Goffinet, D. Heslow, J. Launay, Q. Malartic, B. Noune, B. Pannier and G. Penedo, Falcon-40B: an open large language model with state-of-the-art performance (2023).