Neuro-Symbolic methods for Trustworthy AI: a systematic review with a focus on interpretability

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: msarker@bowiestate.edu

Abstract.
Recent advances in Artificial Intelligence (AI) and especially in deep learning have manifested an increasing concern in trustworthiness. 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, particularly for interpretability. 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 of the recent articles that achieve better interpretability thanks to NeSy systems, while also considering contributions in a broader sense, such as safety, fairness, and privacy. We also did a categorization of the existing contributions along several key dimensions related to the symbolic structures they are exploiting, and the type of interpretability they provide.

Keywords: Neuro-Symbolic, Trustworthiness, 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. 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, and is now a major research focus. 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 interpret 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].

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

Neuro-Symbolic AI (NeSy), which combines Machine Learning (ML) with mechanisms related to Knowledge Discovery and Data Mining (KDD), seeks to integrate neural networks with symbolic processing techniques. This field attracts interest from two distinct perspectives [2]. From a cognitive science perspective, while human brains exhibit connectionist characteristics like neural networks, they are also able to process complex symbolic structures. This capability is believed to play a crucial role in the superiority of human intelligence over other animals. Additionally, from a conceptual perspective, it appears that symbolic and neural approaches complement each other, each with their 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 trustworthiness and especially interpretability. Indeed, one of the main criticism toward the current neural models is their lack of transparency, but symbolic systems do not have this issue. Therefore, in tasks where the state-of-the-art is largely consisting of neural methods, developing a neuro-symbolic approach offers the opportunity to exploit the interpretability that the symbolic aspect provides.

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. The survey methodology was designed to encompass methods that could be considered as neuro-symbolic even-though they were not labeled as such; it is thus presenting a more representative overview and highlighting the amount of existing work that can be classified as neuro-symbolic. We considered different trustworthiness dimensions: privacy, fairness, safety, or interpretability. However, all but two of these papers focused on interpretability, thus interpretability became a primary focus of this work. These papers 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. We also reviewed papers dealing with interpretability based on the symbolic structures used. This review provides an overview of the current trends in this domain, highlighting areas that have been thoroughly explored, revealing common challenges and designs across multiple domains and pinpointing promising directions for future research.

The paper is organized as follows. Section 2 provides an extensive background of the core concepts involved in this survey, as well as grounding for our taxonomy and the presentation of related works. Section 3 presents the survey's methodology, its framing and some observations on the papers found. Section 4 develops on the different types of symbolic knowledge used by the selected papers and presents each paper's contribution. Section 5 analyzes the different types of interpretability provided by these papers. Section 6 presents additional learnings from our review and further discussions.

2. Background

2.1. History of Neuro-Symbolic AI

The genesis of Neuro-Symbolic (NeSy) research is deeply intertwined with the history of Artificial Intelligence (AI), its roots arguably dating back to a seminal 1943 paper by McCulloch and Pitts [3]. This pioneering work used 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, while 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 to create problem solving systems [4]. However, the field encountered unexpected hurdles, leading to the infamous "AI winter" of the 1980s, marked by a significant decline in AI interest and funding [5]. Despite this setback, research on 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 new 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, limited reasoning capabilities, and inherent biases.

NeSy emerged as a promising avenue to address these challenges. Although its conceptual roots span 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). Although neural networks have shown impressive performances, 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]. Furthermore, 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 multimodal processing and applications [7].

2.2. 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 on for high-stakes decisions with minimal or no supervision. Although this certainly includes performance, as a high-performing system is a prerequisite for trustworthiness, in the AI field, trustworthiness encompasses several additional dimensions: *interpretability*, *fairness*, *robustness*, *privacy*, and *safety* [8–10].

Fairness focuses on ensuring that AI models do not harbor biases that could lead to discrimination against certain groups [11]. This is especially pertinent in AI applications involving the classification of people, 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 input [13]. Addressing these biases to ensure fairness towards all groups is a critical concern.

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

Robustness is about 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. The susceptibility of deep neural networks to adversarial attacks is particularly concerning, when subtle manipulations of input data can lead to incorrect interpretations by the AI, despite being obvious to humans. *Since robustness is closely intertwined with performance, it is often unclear 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 the specification of objectives, oversights in the learning process, or other implementation mistakes. As AI systems are increasingly deployed in complex environments with real stakes, 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.

2.3. Background on Interpretability

Interpretability is the most extensively addressed aspect of trustworthiness, experiencing exponential growth as a research domain [14, 15]. There is little consensus on the precise definition of interpretability, but it can be broadly defined as the extent to which a system's operations can be understood by users [1, 16]. This includes access to mechanisms or reasoning that underpin the system's predictions. Simpler systems are naturally more interpretable, which is why this was not 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, for societal acceptance

as well as regulatory compliance; indeed, both United States and the European Union have imposed a right to explanation for consumers [14]. 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. Although many papers use explainability and interpretability interchangeably, some argue that explainability is a stronger concept than interpretability [14, 17, 18]. Since the frontiers between these two concepts is quite vague, our review is based on the terminology used by the authors of the papers, which may not always align with this distinction; thus explainable and interpretable should be understood as interchangeable in our paper. Generally, interpretability is self-assessed by researchers, leading to calls for more rigorous taxonomies and evaluations [16, 19–21]. It is also important to note that explainability is not the "silver bullet" for AI trustworthiness. Studies have shown that while explainability can enhance AI collaboration with novices, it does not necessarily do so with experts [22]; a combination of AI and human decision-making can be quicker but less accurate when AI provides explanations [20]; and there is a risk that explanations, even if not particularly useful, can unduly increase public acceptance, leading to over-reliance on AI [23, 24].

Interpretability in AI systems has been tackled through a variety of methods, which can usually be divided into two categories depending on what kind of interpretability they provide: either ***ante-hoc*** or ***post-hoc*** [15, 19, 21]. ***Ante-hoc*** *explainability* is a characteristic of systems which are inherently designed to be easily interpretable from the inside (also known as *self-explainable* systems). These systems are structured in order to make their internal processes straightforward and clear. ***Post-hoc*** *explainability* is a characteristic of systems that are not inherently transparent but present an additional layer that attempts to give more insights about the underlying decision process. 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 [25], who argues against the use of such explainability, suggesting that it can be misleading. Conversely, Gilpin et al. [18] argued that when using post-hoc explanations, it is crucial to clearly inform users about their potential limitations. There are also approaches that fall somewhere between these two extremes [26, 27]. These methods aim to train systems in a way that makes their decision-making processes easier to interpret, without fundamentally altering their core structure.

Interpretability systems can also be categorized based on the scope of explanations, they can range from ***local*** to ***global*** [15, 19, 21]: ***Local*** *explanations* are explanations tailored to individual instances, providing insights on specific decisions or similar cases. A well-known example of a local explanation method is LIME [28], which is designed to offer explanations for specific data points. ***Global*** *explanations* aim to shed light on the system's behavior as a whole, irrespective of individual inputs. Some methods [29, 30] provide explanations for a specific category of inputs, allowing 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.

2.4. 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. [31] have even suggested that symbols are essential for effective communication between humans and AI systems. Although 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 implicit 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. [14] 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 integrate 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.

2.5. 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. [32] addresses recent techniques for enhancing AI trustworthiness. This work examines trustworthiness over six dimensions: explainability, robustness, accountability /auditability, privacy, fairness, and environmental well-being. Another notable review in the field of trustworthy Machine Learning was conducted by Serban et al. [33], providing valuable insights into methods to foster trust in AI systems.

Interpretability methods in AI have received considerable attention, with numerous reviews dedicated to this topic. In the latest reviews, the focus is usually expressed with the word *explainability* (XAI), but as discussed in Section 2.3, the core idea is the same as *interpretability*. For instance, Speith et al. [15] conducted an analysis of various taxonomies used to categorize 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. They argued that the choice of taxonomy should align with the user's needs and proposed a unified taxonomy to guide users. Similarly, the very comprehensive review by Barredo Arrieta et al. [21] analyzed and synthesized the existing taxonomies of XAI, and proposed to assess explainability based on the targeted audience. They also proposed both a review of existing transparent (ie. self-interpretable) methods and a review of post-hoc explainability methods. Lastly, they extended their review to what they call *"Responsible AI"*, a concept roughly equivalent to Trustworthy AI. Another relevant review by Weller [23] explored the concept of transparency, putting it in perspective with the different stakeholders of AI. This consideration of the different stakeholders is also of key importance in the frameworks of Kasizadeh [1] and Langer et al. [34]. Vilone et al. [35] 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.

Reviews specifically focusing on NeSy learning have also been published [4, 7, 36–38]. Sarker et al. [36] 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. [7] presented a more conceptual review of the neural-symbolic field, discussing its foundations, current applications, and future challenges. Berlot-Attwell [37] explored the use of NeSy AI in Visual Question Answering (VQA), while Hamilton et al. [38] 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. Wang et al. [4] conducted a systematic overview of recent advances in neuro-symbolic computing and described a taxonomy in four dimensions, inspired by Bader and Hitzler [39].

While we kept in mind the learnings from the existing reviews, we believed that the intersection between NeSy and interpretability was an interesting novel area to review. To our knowledge, this paper may be the first to review Neuro-Symbolic methods specifically through the lens of Trustworthy AI.

3. Survey

3.1. Methodology

The aim of this survey was 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 2022 (latest publications at the time of writing). While we are aware of emerging venues such as the NeSy AI Journal and those hosted by IOS Press and Sage, many of them were either very recent or not yet fully established at the time of our review. Additionally, although several well-regarded AI journals exist, their primary focus did not align specifically with the neuro-symbolic AI domain. As a result, we focused our literature review on top-tier AI conference proceedings, where substantial and timely research in this area was more readily accessible. We selected

papers from the following conferences: *NeurIPS*, *AAAI*, *IJCAI*, *IJCL*, *ICML*, *NeSy*, *AACM FAccT*, and *KDD*. The volume of papers presented at these conferences, exceeding 10,000 over the last two years, required 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. Using the *dblp* database, papers were intially filtered 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*. Search-in-page tools were then used 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)*, *safe* and *understandab(le)*. Papers meeting these criteria were examined in more detail to assess their relevance to our focus.

Additionally, to ensure that we did not overlook papers that explicitly mentioned the use of NeSy 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* (and thus every derived terms such as *symbolic*) 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 research and trustworthiness in AI.

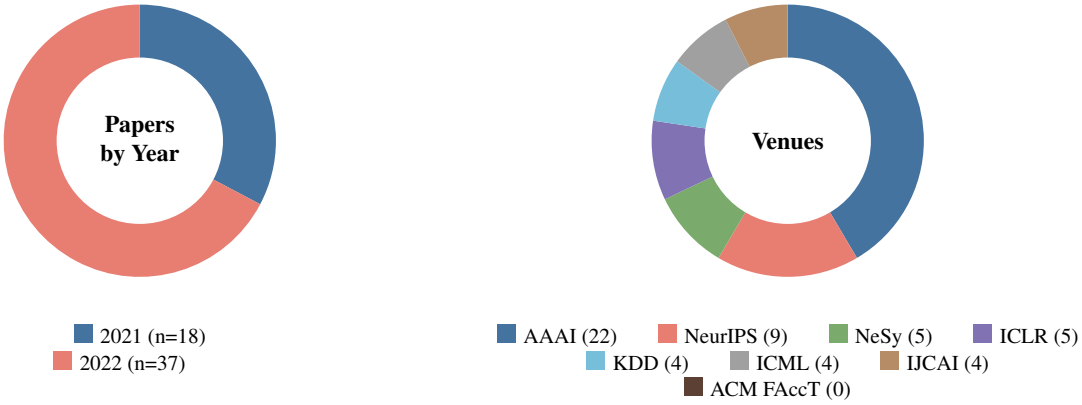

Fig. 1. Distribution of selected papers.

3.2. *Framing the survey*

Determining whether a paper's approach qualifies as Neuro-Symbolic 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 graph that was neither used nor executed in the system, we did not consider the symbolic role to be sufficiently integrated in the framework for it to be considered as 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 by-product rather than a primary focus, without substantial discussion or emphasis. To maintain the relevance and specificity of our survey, we chose to include only the 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.

3.3. Some Statistics

Our comprehensive review yielded a total of 54 papers that employed neuro-symbolic 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 [40] and another on safety [41]), focused 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 to address trustworthiness concerns beyond interpretability. Another observation is that of a significant increase in relevant publications in 2022, with 37 out of the 54 papers coming from this year alone (Figure 1), indicating a growing interest and expansion in this domain. The distribution of these papers across various conferences reveals that AAAI is the predominant venue for this type of research. We also noted that no papers from ACM FAccT ended up in our survey, despite the fact that this conference specifically focuses on trustworthiness issues. This is because our keyword search found very few matches in FAccT papers, and the only papers that matched were not proposing a neuro-symbolic approach in our view. This was quite surprising, but it is worth mentioning that this is also a conference with quite few papers compared to the other conferences considered here.

3.4. Applications of NeSy Systems

While exploring interpretability contributions of the NeSy systems, we found that they were being used for different applications. Many of the proposed systems were working with visual data: either image classification [26, 30, 42, 43], action recognition in videos [44], agent communication about images [45], handwritten mathematical expression recognition [46], visual relation detection [47], or visual reasoning [48]. Equally many of the systems dealt with natural language settings: fake news detection [49–51], question answering [52, 53], unspecified NLP [29], text classification [40], commonsense reasoning [54], medical diagnosis through dialogue [55], text fiction tasks [56], or news recommendation [57]. A few applications were entirely based on graphs: knowledge graph completion [58, 59], query answering on knowledge graphs [60], graph classification [61], or imitating algorithms on graphs [62]. Some researchers worked in settings where an agent has to make different decisions (often trained with reinforcement learning) [63–66]. In some cases, the methods were explicitly suited for multiple settings [46, 67]. Lastly, many other unique settings were explored: adaptive management [68], time series analysis [69], congestion control [70], safe execution of programs [41], or computer algebra [71]. The wide range of applications shows how versatile NeSy methods can be.

4. Analysis based on the type of symbolic knowledge

4.1. The different symbolic structures used

In our categorization of the papers, we found that 16 of them presented ***rule-learning*** approaches [72–87]. These papers typically utilize deep learning to generate logic rules or decision trees for classification purposes. In these instances, machine learning techniques allow the creation of a symbolic model, which offers 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. We identified that these systems could be broadly categorized into three types based on the symbolic data structures they manipulate: *logic*, *graphs*, and *other structures*. This classification, depicted in Figure 2, provides insight into the varied approaches within the NeSy field, highlighting the diversity of methods being explored to improve trustworthiness in AI systems.

Logic, as used in various papers [26, 30, 40, 43, 50, 54, 59, 61, 62, 67, 69, 70, 88], typically involve logic propositions, often in the form of logic rules (e.g., *precondition* $\rightarrow$ *class*). This approach usually 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 [56–58, 60], but the category also includes other types of graphs [46, 49, 51–53, 68], such as scene graphs, proof graphs, or Abstract Meaning Representation

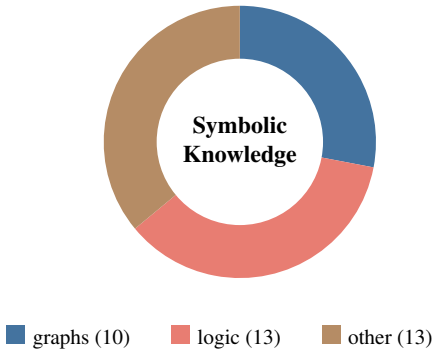

Fig. 2. Form of the symbolic knowledge *(in papers that do not fall into the rule-learning category)*

(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 [27, 29, 42, 44, 45, 47, 48, 55, 63–66, 71]. This includes, for example, symbolic descriptions of objects or symbolic programming languages. This category is diverse and encompasses a wide range of approaches in which symbolic representations take on various forms.

4.2. Description of the reviewed papers

Interestingly, each of these three categories—logic structures, graphs, and other structures—encompasses a similar number of papers. 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. *Note that we chose not to elaborate about rule-learning papers [72–87] in the following to maintain appropriate scope and manuscript length.*

4.2.1. Approaches using logic

A common approach for interpretability is to extract symbolic rules that explain the system's prediction. This can often involve modifying the system's structure so that the rule extraction from it can be more feasible and faithful. For instance, Barbiero et al. [67] proposed a new approach in which the classifier is designed in a way that allows the extraction of logic rules to explain its predictions. This approach can be related to Lee et al.'s framework [88], which upgrades a deep model into a self-explainable version by naturally integrating human priors and rule generation into its predictions. Acting on the training step, Sharan et al. [70] proposed a method to train a deep model, then extracts from it symbolic rules. Similarly, Kasioumi et al. [26] proposed a new learning method (Elite BackProb) which promotes activation sparsity of the filters of a convolutional neural network, so that a rule extraction algorithm can be used to approximate its predictions. In the graph processing domain, Himmelburger et al. [61] made a framework which extract rules for post-hoc explanation of graph neural networks (GNN). Georgiev et al. [62] proposed concept-bottleneck GNNs, which are variants of GNNs with a new readout which allows the production of explanations in propositional logic based on inferred concepts. Cucala et al. [59] proposed a new class of knowledge graphs transformations that are always equivalent to the application of symbolic rules. Rajapaksha and Bergmeir [69] proposed a model to produce rule-based explanations of a black-box Global Forecasting Model on several time series.

Other approaches used logic in original ways, usually involving it in the system's decision to make it more interpretable. For example, Chen et al. [50] proposed an approach that decomposes texts into phrases and uses aggregation logic to classify them as fake or not in an interpretable way. Yao et al.'s framework [40] parses advices on what a language model is wrongfully using for its prediction into First Order Logic (FOL) and use it to refine the weights of the language model. In this case the main goal is about increasing fairness and not interpretability. Ribeiro and Leite's framework [30] maps a neural network's internal states to concepts from an ontology, making it possible to build explanations from these concepts. Kalyanpur et al. [54] proposed a novel reasoner that combines symbolic reasoning with statistical functions for fuzzy unification and dynamic rule generation. In the image classification

domain, An et al. [43] proposed a novel rule-guided method called dynamic ablation to provide explanations as well as visual highlights.

4.2.2. Approaches using graphs

Among the works using graphs, a few approaches share the common characteristic of using knowledge graphs. For instance, Zha et al. [58] proposed a method for knowledge graph completion that outputs a pattern in the graph to explain the predictions, using BERT. Zhu et al. [60] developed an approach that converts logical queries into circuits including graph neural networks to answer them based on a knowledge graph. Liu et al. [57] proposed a new method for news recommendation: small anchor graphs are generated via reinforcement learning so that the similarity of two articles can be estimated by computing the number of paths connecting the two anchor graphs. Peng et al. [56] designed a reinforcement learning agent that uses a knowledge graph to represent its belief about the world alongside an attention mechanism to be able to explain its reasoning.

Various approaches also used different types of graphs in original ways. For instance, Ferrer-Mestres et al. [68] proposed an approach to extract policies of a fixed length from policies or arbitrary lengths so that they are small enough to be interpretable, in the Mixed Observability Markov Decision Process Setting (policies are represented as graphs). Wu et al.'s framework [46] decomposes images of mathematical formulas into graphs to make the process of inferring the formula more interpretable. Zhong et al. [53] proposed a method that produces hybrid chains (mixing text and table data) and reason on those with a transformer to provide an answer, in a question answering (QA) setting. For the same task, Deng et al. [52] proposed a method that parses questions into AMR (abstract meaning representation) graphs and reason on those graphs to answer the questions. For the task of fake-news identification, Jin et al. [51] took inspiration from human's information-processing model to make a model that builds claim-evidence graphs to identify fake news. For a similar task but focusing on the propagation network of fake-news, Yang et al. [49] designed a framework that reveals which subgraphs of the news propagation network are the most important in a model's decision process.

4.2.3. Approaches using other forms of symbolic knowledge

Many papers used various forms of symbolic data, usually specific to their application. Some of them worked with autonomous agents, often trained with reinforcement learning agents. For instance, Sreedharan et al. [66] proposed a method to provide contrastive explanations with user-specified concepts in sequential decision-making settings, by building partial symbolic models of a local approximation of the task. In a similar setting, Jin et al. [65] developed a framework to learn action models and symbolic options with a symbolic planner, using reinforcement learning. Also considering agents trained with reinforcement learning, Finkelstein et al. [64] designed a protocol to apply transformations to the environment model of an autonomous agent in order to produce textual explanations of it. Targeting a broader range of models, Verma et al. [63] proposed a new approach based on query answering to estimate a black-box autonomous agent as an interpretable relational model.

In the medical domain, Jang et al. [42] proposed an approach that extracts symbolic representations from images and rules on these representations to diagnose some diabetes. Another paper in the medical domain, by Liu et al. [55], presented a method for medical diagnosis, that uses a Bayesian Network as well as conditional probability and mutual information matrices to direct an inquiry of the symptoms in order to identify a disease.

In the visual domain, Chen et al. [47] proposed a method that combines deep learning with analogical learning on visual relation detection, using object information and spatial information between objects, so that the relation identification relies on an interpretable algorithm. For dynamic visual reasoning in videos, Ding et al. [48] proposed a method that identifies objects and related physics concepts such as speed, then gives them as input to a physical simulator to predict what will happen next. Hua et al. [44] developed a method that consists of decomposing videos into object-relation chains, which allows both the classification and the production of explanations based on this representation.

Lastly, we observed a few original papers which are either generic or have a domain of focus which is not shared with other papers in this category. For instance, Geiger et al. [27] proposed a training method that allows the alignment of the neural network to a high-level causal model. Dessi et al. [45] presented a protocol to train two deep neural networks with a way of communicating using symbols, which is partially interpretable. Zhang et al.'s framework [29] extracts from a deep Natural Language Processing (NLP) model the interaction between the words

that influenced the embedding, and outputs a tree structure. Peng et al. [71] proposed a new framework for symbolic computation that decomposes computations in fundamental transformations, performed by deep models.

4.3. Synthesis and Observations

Overall, across the three categories of symbolic knowledge, several common trends and challenges emerge. First, the diversity of symbolic structures used across the three categories reflects the versatility of symbolic representations in AI, and suggests that the choice of symbolic structure is often driven by the specific application domain. For instance, graphs are more prevalent in natural language processing tasks, while logic rules are more commonly used in image classification and graph processing. Second, a notable observation is that many of the papers in these categories are domain-specific, with a significant number focusing on natural language processing, visual reasoning, and autonomous agents. This suggests that the development of NeSy methods for trustworthiness is still largely driven by specific application needs, rather than by a unified theoretical framework. Finally, while the majority of papers across all three categories focus on interpretability, the approaches differ in the type of interpretability they provide. Logic-based approaches tend to provide more formal and precise explanations, while graph-based approaches often provide more intuitive and visual explanations. The "other" category, being more diverse, encompasses a wider range of interpretability types, from contrastive explanations to causal models. These observations suggest that, despite the variety of symbolic structures and application domains, there is a common underlying goal of bridging the gap between the high performance of neural models and the transparency of symbolic systems.

5. Classification based on the types of interpretability

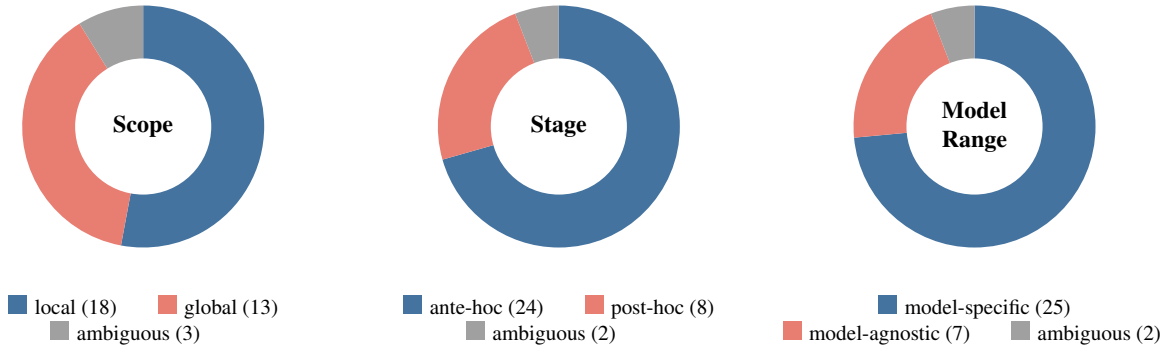

Fig. 3. Distribution of the different types of interpretability contributions.

Since interpretability represents the central theme of the majority of the collected studies, we systematically categorize these works to better analyze prevailing methodological trends. Specifically, the papers are organized along three widely recognized dimensions in interpretability research: (1) the **scope** of explainability, (2) the **stage** at which the interpretability method is applied, and (3) the degree of dependence on the underlying model architecture (**model range**). This taxonomy provides a structured view of how interpretability techniques are designed and deployed across the literature. The resulting distribution of studies across these dimensions is illustrated in Figure 3, while the detailed classification of individual papers is summarized in Table 1. These dimensions are frequently used to analyze papers in this field [15, 19, 21] and the description of those concepts is mentioned in Section 2.3. *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 of our taxonomy concerns the scope of explainability, which distinguishes between explanations that focus on individual predictions and those that describe the behavior of the model as a whole. Accordingly,

interpretability methods in the literature are commonly categorized into local explanations, which explain a specific prediction or instance, and global explanations, which aim to capture the overall decision-making logic of the model. ***Local*** explanations are specific to a given input, providing insights into why a particular decision was made. In this case, the explanations are different for each individual processing; in other words, running the model several times on different inputs will give different explanations. On the other hand, ***global*** explanations offer a broader understanding, characterizing the behavior of the entire model. In between, an intermediate *"cohort"* scope is sometimes considered, for methods applicable to a subset of inputs rather than just one[1]. Our review found a relatively balanced distribution: 18 papers (53%) focus on local explanations, 13 papers (38%) on global explanations, and 3 papers (9%) with an ambiguous scope. This shows that both strategies are well studied in the literature.

Regarding the **stage** of explanation, methods can be categorized as either ***ante-hoc*** (also known as *self-explainable*) or ***post-hoc***. ***Ante-hoc*** interpretability contributions are frameworks designed to be inherently explainable, usually by making part of the decision process structured and comprehensible by humans. On the other hand, ***post-hoc*** methods generate explanations after or in parallel to the decision, 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 fundamental difference of this in contrast with *ante-hoc* interpretability is that the explanation does not necessarily align with the real reasons why a particular output was produced by the framework. This means that it wouldn't help to uncover potential bias or imperfections in the reasoning process, and might even hide problems in the decision process with justifications *a posteriory*. Therefore, it is questionable how such papers actually contribute to Trustworthiness which was our initial theme. In our study, we found that most of the papers covered provided ante-hoc interpretability contributions (24 ; 71%), and only 8 of them (24%) provided post-hoc interpretability contributions. Our survey found no clear correlation between the scope of explanations and the stage at which the method is applied, indicating a wide range of approaches addressing interpretability in AI systems.

The third dimension (**model range**) concerns whether the interpretability method is ***model-agnostic*** or ***model-specific***. ***Model-agnostic*** methods can be applied universally across different models, while ***model-specific*** methods are tailored to a particular model. In this dimension, we observed almost the same distribution than for the previous categorization: 25 papers (74%) proposed a model-specific interpretability contribution, while 7 of them (20%) proposed a model agnostic contribution. In fact, this similarity is not due to chance, since we can see in table 1 that all ante-hoc explainability contributions fall into the model-specific category. This makes sense, because providing ante-hoc interpretability requires designing an inherently explainable method, so they are naturally model-specific. On the contrary, post-hoc explainability methods have the flexibility to be model-agnostic, and we observed that all but one of the methods in the post-hoc category were model-agnostic. An interesting exception we noticed is the work by Seungeon Lee [88], which involved modifying the final layer and training process of a deep model to enhance explainability.

From this classification, we derive several key insights. First, local explanations slightly outnumber global ones, suggesting a focus on instance-level interpretability in neuro-symbolic research. Second, ante-hoc methods dominate, reflecting the field's emphasis on intrinsic interpretability rather than post-hoc explanations. Finally, the prevalence of model-specific methods indicates that neuro-symbolic interpretability often requires custom design rather than generic tools.

6. Discussion

6.1. Lack of applications to Fairness, Privacy and Safety

The primary objective of this review was to examine how Neuro-Symbolic (NeSy) systems are applied to address different dimensions of AI trustworthiness. As expected, interpretability emerged as the dominant research focus in the literature. However, the limited number of studies addressing other trustworthiness aspects—particularly fairness and privacy—was notable. Although some neuro-symbolic approaches targeting safety, privacy, or fairness

[1] labeled in the figure as "ambiguous" for consistency with other dimensions

Table 1
Classification of papers by interpretability dimensions (see Section 5)

Scope (Local vs Global)	**Local:** [42–44, 46, 50, 51, 53, 54, 56, 57, 60, 64–69, 71]	**Global:** [26, 27, 48, 50, 52, 55, 58, 59, 61–63, 70, 76]	**Ambiguous:** [29, 30, 45]
Stage (Ante vs Post)	**Ante-hoc:** [42, 44–48, 50–60, 62, 65, 67, 70, 71, 76, 84, 88]	**Post-hoc:** [29, 30, 43, 61, 63, 64, 66, 69]	**Ambiguous:** [26, 27]
Type (Spec. vs Agn.)	**Model-specific:** [29, 30, 42, 44–48, 50–60, 62, 65, 67, 70, 71, 76, 84]	**Model-agnostic:** [43, 61, 63, 64, 66, 69, 88]	**Ambiguous:**[27, 29]

may exist, their absence from the prominent conferences/journals (from where we collected the papers) suggests that these topics currently have limited visibility or adoption within the NeSy research landscape.

This suggests that there may be unexploited potential. 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. Regarding safety, we did find one paper [41], so there seems to be at least some potential. The scarcity of NeSy approaches to safety could be attributed to the fact that safety is often seen as a broad concept with a lot of overlap with other dimensions, such as robustness for instance, and it is this dimension which appears as a clear goal in the publications. Safety, which is the study of worse-case scenarios and respect of critical constraints, encompasses considerations that are very specific to the target applications. Looking for the keyword "safe" in AAAI accepted papers, we observed that it was much less common than what we could find for "interpretab(le)" and "explainab(le)" combined, and that it appeared mostly in publications about reinforcement learning settings. We could hypothesize that neuro-symbolic approaches are less popular in these settings, and that it contributes to the rarity of NeSy approaches to safety.

Considering fairness, there seems to be untapped potential for NeSy integration. In addition to the paper we mentioned previously [40], we found another work by Wang et al. [89], which 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 the integration of fairness constraints could be facilitated by NeSy models. This suggests that further investigation into NeSy's potential to address fairness in AI should be pursued. Moreover, there is a close link between interpretability and fairness, since having more understanding of a system's decision could help to detect potential biases, as pointed out by Barredo Arrieta et al. [21]. This idea is supported by some works [90, 91] that address both explainability and fairness simultaneously. However, post-hoc explainability introduces a new risk of "fair-washing", which is to give in appearance fair explanations to decisions that were taken because of biases [92]. Overall, since interpretability and fairness are related and NeSy designs have a high potential for interpretability, we believe that they have a high potential for fairness as well, which definitely requires further research.

6.2. Lack of grounding based on common taxonomies

Another insight 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. Although review papers like those by Sarker et al. [36] and Wang et al. [4] proposed 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. Our survey regrouped papers of different NeSy categories according to Wang et al. [4], and these categories regrouped papers of different application domains, suggesting that several researchers could face the same challenges without awareness of the insights from other fields. However, we could not always be sure of each framework's category. More clear grounding on taxonomies would facilitate the identification and comparison of works with similar methodologies, regardless of their specific applications.

Regarding the interpretability, there are numerous existing works on the taxonomies (see Section 2), but the papers are too rarely providing clear grounding of their work on those. For instance, how consistent are the explanations with the actual decision-making is not often actually assessed; one usually needs to understand in depth the proposed method to evaluate the explanations consistently, despite this issue being of key importance. Another takeaway from our review is that it is very hard to quantify the degree of explainability and thus compare the different methods. Although some papers provided user studies [49, 55, 58, 66, 88], it is far from a universal practice, and user studies are not standardized. More systematic assessment, and standardised metrics, would make it possible to deduce actionable insights from the comparison of the different methods.

6.3. Common Challenges

One of the main challenges faced by interpretability methods is the trade-off between interpretability and performance. The reason for the domination of "black-box" deep models over the state-of-the-art in many domains is that they have shown empirically to be the best performing. While post-hoc explainability approaches keep the neural model untouched and thus do not alter their performance, self-interpretable frameworks are designed specifically for interpretability, thus their performance may not be optimal. However, as it was pointed out in earlier studies [17, 25], a trade-off between interpretability and performance is not systematic. Indeed, we found that a large part of our surveyed papers claimed new state-of-the-art results for their task, and a few others claimed competitive results with SOTA (most of the time neural approaches). Even if some of the papers showed performances below the SOTA or did not provide comparison with non-interpretable approaches, we can definitely say that neuro-symbolic approaches have the potential to be the best performing while providing interpretability, in a lot of domains. Therefore, we strongly recommend the exploration of NeSy strategies.

A common challenge for NeSy approaches in general is to design the interface between neural components and symbolic parts. When the symbolic knowledge is altering the training of the neural components (often through the loss), finding the right integration and the right weight is very challenging. When neural components need to output intermediate symbolic parts, the model cannot be trained end-to-end, and a main challenge is to choose the right symbolic space, as well as to get the model to provide outputs in this space. When the symbolic knowledge is restraining the output space of neural networks or acting directly on the inference process, a main challenge is to ensure good design so that the learning process is not hampered. On top of these difficulties, ensuring interpretability should be considered when designing the framework, since NeSy systems are not always interpretable. This often implies giving a strong enough role to the symbolic structures, especially at the steps leading to the final outputs. The actual explainability provided by the symbolic representations is also to be considered. All of these design challenges are common to methods from different domains using similar methodologies, therefore leveraging insights from designs in other domains is essential when approaching a new task.

Another challenge faced by NeSy systems is that of scalability. Regarding the symbolic structures, this means scaling the symbolic space to increase the potential expressivity and cover more cases, involving richer knowledge. However, more expressive symbolic spaces make the symbolic integration more difficult. For instance, many methods build rules in Propositional Logic, which lacks the expressivity of First-Order Logic. It is often difficult to scale the methods to First-Order Logic, one of the reasons being that it introduces a combinatorial quantity of possible formulas and possible reasoning patterns. For the neural components, the problem of scalability is mainly about data. Indeed, many NeSy approaches require datasets with additional structural information, or datasets of intermediate symbolic representations. It is hard to find such data in large quantities, while neural models often require a lot of data with a diverse enough distribution to be reliable and robust. Future research should explore ways to enhance the scalability of NeSy approaches.

7. Conclusion

This research work provides 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 current NeSy methods for trustworthiness is centered around improving interpretability. By converging the fields of AI trustworthiness and NeSy integration, this study proposed a new unified analysis of these two intertwined domains.

The papers included in our survey were reviewed based on the symbolic structures they were exploiting. They were also systematically categorized on the basis of the scope, stage, and adaptability of the interpretability methods they developed. A key insight from our study is the recognition of the immense potential NeSy integration holds for interpretability. This potential is not restrained to any specific domain or application, indicating a broad and versatile utility of NeSy approaches.

However, this study also highlights a significant 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 works 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. This study also uncovered a lack of grounding on existing taxonomies, and the lack of standardized assessment of interpretability.

In conclusion, this review provides insights on how the emerging research trend of improving trustworthiness via NeSy methods can be analyzed and structured. It should facilitate a comprehensive understanding of the field, and open 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] T.L. Johnson, N.N. Johnson, D. McCurdy and M.S. Olajide, Facial recognition systems in policing and racial disparities in arrests, *Government Information Quarterly* **39**(4) (2022), 101753. doi:https://doi.org/10.1016/j.giq.2022.101753. https://www.sciencedirect.com/science/article/pii/S0740624X22000892.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila and F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion* **58** (2020), 82–115–. doi:10.1016/j.inffus.2019.12.012.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] M.T. Ribeiro, S. Singh and C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, B. Krishnapuram, M. Shah, A.J. Smola, C.C. Aggarwal, D. Shen and R. Rastogi, eds, ACM, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing and K. Baum, What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research, *Artificial Intelligence* **296** (2021), 103473. doi:10.1016/j.artint.2021.103473.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] S. Bader and P. Hitzler, Dimensions of Neural-symbolic Integration - A Structured Survey, in: *We Will Show Them! Essays in Honour of Dov Gabbay, Volume One*, S.N. Artëmov, H. Barringer, A.S. d'Avila Garcez, L.C. Lamb and J. Woods, eds, College Publications, 2005, pp. 167–194.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] E. Soares and P. Angelov, Fair-by-design explainable models for prediction of recidivism, arXiv, 2019. doi:10.48550/ARXIV.1910.02043. https://arxiv.org/abs/1910.02043.

[91] Y. Ahn and Y. Lin, FairSight: Visual Analytics for Fairness in Decision Making, *IEEE Trans. Vis. Comput. Graph.* **26**(1) (2020), 1086–1095. doi:10.1109/TVCG.2019.2934262.

[92] U. Aïvodji, H. Arai, S. Gambs and S. Hara, Characterizing the risk of fairwashing, in: *Advances in Neural Information Processing Systems*, Vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang and J.W. Vaughan, eds, Curran Associates, Inc., 2021, pp. 14822–14834. https://proceedings.neurips.cc/paper_files/paper/2021/file/7caf5e22ea3eb8175ab518429c8589a4-Paper.pdf.

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

[94] 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.

[95] 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.

[96] 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.

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

[98] 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.

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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).