Proceedings of the of the 18th Conference on Computer Science and Intelligence Systems pp. 1–10
DOI: 10.15439/2023F6019
ISSN 2300-5963 ACSIS, Vol. 35

Measuring Trustworthiness in Neuro-Symbolic Integration

Andrea Agiollo, Andrea Omicini
0000-0003-0531-1978, 0000-0002-6655-3869
ALMA MATER STUDIORUM – Università di Bologna, Italy
Email: {andrea.agiollo, andrea.omicini}@unibo.it

***Abstract*—*Neuro-symbolic* integration of symbolic and subsymbolic techniques represents a fast-growing AI trend aimed at mitigating the issues of neural networks in terms of decision processes, reasoning, and interpretability. Several state-of-the-art neuro-symbolic approaches aim at improving performance, most of them focusing on proving their effectiveness in terms of raw predictive performance and/or reasoning capabilities. Meanwhile, few efforts have been devoted to increasing model trustworthiness, interpretability, and efficiency—mostly due to the complexity of measuring effectively improvements in terms of trustworthiness and interpretability. This is why here we analyse and discuss the need for ad-hoc *trustworthiness metrics* for neuro-symbolic techniques. We focus on two popular paradigms mixing subsymbolic computation and symbolic knowledge, namely: *(i) symbolic knowledge extraction* (SKE), aimed at mapping subsymbolic models into human-interpretable knowledge bases; and *(ii) symbolic knowledge injection* (SKI), aimed at forcing subsymbolic models to adhere to a given symbolic knowledge. We first emphasise the need for assessing neuro-symbolic approaches from a trustworthiness perspective, highlighting the research challenges linked with this evaluation and the need for ad-hoc trust definitions. Then we summarise recent developments in SKE and SKI metrics focusing specifically on several trustworthiness pillars such as interpretability, efficiency, and robustness of neuro-symbolic methods. Finally, we highlight open research opportunities towards reliable and flexible trustworthiness metrics for neuro-symbolic integration.**

I. INTRODUCTION

A GROWING number of critical applications are being developed that rely on artificial intelligence (AI) solutions—mostly, on machine and deep learning (ML, DL), more specifically. In this realm, the most popular trend is by far the engineering of intelligent computational systems where hard-to-code tasks are automatically learned from data—promoting a data-driven problem-solving approach. Tasks that can be learned this way range from image [1], [2] to text processing [3], [4], stepping through graph learning [5], [6], [7] and time series forecasting [8], [9], among the many others. The popularity of (semi-)autonomous AI systems largely depends on their ability of outperforming humans in some specific tasks. Yet, AI agents – and especially ML agents – cannot be really trusted by humans, for the obscurity of their data processing and decision making pipeline, and for their limited interaction with human users as well. In the recent past, this lack of trustworthiness jumped to the news due to some AI systems' behaviour harming humans—such as chatbots suggesting deleterious practice[1] and facial-recognition technology recognising innocents as criminals.[2] Therefore, the need to assess the level of trustworthiness of AI system before its deployment it is nowadays apparent to all parties involved in the development of AI solutions. Targeting this need, the European Union (EU) has recently released the Ethics Guidelines for Trustworthy AI[3] as a part of its AI strategy.

While representing a fundamental stepping stone in the definition of AI trustworthiness, these ethics guidelines apparently focus on popular ML agents solutions in their definition process. Indeed, most trust requirements are clearly linked with the black-box nature of ML and DL solutions—such as the need for transparency, explanations, human interaction, and many others. However, AI is not just ML/DL, so AI systems are much more than ML/DL systems. Recent research efforts have focused on novel AI paradigms aiming at blending the subsymbolic perspective of ML and DL agents with symbolic AI solutions focusing on high-level symbolic (human-readable) representations of problems, logic, and search: this is where *neuro-symbolic integration systems* (NeSy) stand today. NeSy integrate neural (subsymbolic) and symbolic AI solutions aiming at suitably complementing their strengths and weaknesses, introducing reasoning and cognitive capabilities (the symbolic way) while preserving fast-learning capabilities (the subsymbolic way). The range of NeSy approaches is vastly distant from the AI systems accounted for in the definition of EU trustworthiness pillars, as they leverage symbolic (human-comprehensible) solutions which are in principle trustworthy by design. Therefore, NeSy introduces a further level of complexity in the definition of their trustworthiness value, given by the complex interaction between symbolic and subsymbolic elements. The result is the current lack of suitable definitions of the notion of trustworthiness in terms of NeSy systems.

This is why in this paper we deal with the definition of trustworthiness for NeSy systems, focusing specifically on two broad NeSy categories, namely:

[1] https://edition.cnn.com/2023/06/01/tech/eating-disorder-chatbot/
[2] https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html
[3] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

- *symbolic knowledge injection* (SKI) models—that is, systems featuring symbolic knowledge that can be explicitly provided so that subsymbolic predictions are either computed as a function of it, or made consistent with it;
- *symbolic knowledge extraction (SKE)* models, represented by the set of approaches accepting subsymbolic predictors as input and producing symbolic knowledge as output; the aim for the SKE system is to extract symbolic knowledge reflecting the behaviour of the predictor with high fidelity.

The definition of requirements for trustworthy NeSy systems represents a fundamental step towards their safe adoption. However, requirements definition by itself can not be considered as an exhaustive measure to ensure and calibrate the trustworthiness of NeSy systems. Instead, it is of utmost significance to define NeSy *trustworthiness metrics* that allow to actually measure the level of a system trust, possibly enabling an in-depth analysis of the components raising trust concerns. Whereas a few trustworthiness metrics definition already exist, tackling specific components of NeSy models – such as accuracy, robustness and efficiency –, the vast majority of NeSy most relevant aspects are still unexplored.

This is why in this paper we:

- define how the AI trustworthiness requirements translate to the NeSy realm, analysing in detail each pillar of trust and its implication on NeSy models.
- analyse the available metrics for each of the novel NeSy trust requirements as well as the potential future directions to explore in the analysis of NeSy trust;
- suggest some novel metrics to measure specific NeSy elements, focussing on SKI and SKE as two well-defined broad categories of NeSy models.

This article is organised as follows. Section II presents the transition from trustworthy AI requirements to their corresponding NeSy trust pillars, analysing in depth how NeSy elements impact each requirement. Section III showcases the need for defining trust metrics – rather than requirements – and analyses the complexity of that definition, the reason behind it, and how we propose to tackle it. We then introduce the relevant concepts of SKI and SKE needed to design trust metrics in Section IV, and propose a detailed analysis of available and lacking metrics in Section V. Finally, we conclude and present the future directions in Section VI.

II. From Trustworthy AI to Trustworthy NeSy

As a fundamental step of its AI strategy, the European Union (EU) has defined seven key trustworthiness criteria to meet during the development, deployment, and use of AI systems, namely: *(i) human agency and oversight*, as the need for oversight mechanisms enabling the informed interaction between the AI agent(s) and the human(s) counterpart; *(ii) robustness and safety*, as the need for accuracy, reliability, resilience and security of AI agent(s); *(iii) privacy and data governance*, as the need for ensuring legitimised access to data, while taking into account data quality and integrity; *(iv) transparency*, as the need for providing human users with explanations of the AI agent(s)'s decision process; *(v) diversity, non-discrimination and fairness*, as the need for avoiding unfair bias while enable everyone's access to AI technology; *(vi) environmental and societal well-being*, as the need for sustainability of AI agent(s) and the transition to their environmentally friendly development; *(vii) accountability*, as the need for mechanisms that ensure responsibility and accountability for the behaviour and outcomes of AI systems. The above requirements define a broad umbrella of concepts and means to identify relevant components in the deployment of AI systems and ensure their trustworthiness. However, being designed to be general enough to be applicable to any – or at least as most as possible – AI systems, they are actually too general to be used to define actual metrics to effectively measure every sort of AI systems. Therefore, to make them actually working, a more detailed specification of trustworthiness requirements is needed: in particular, the general EU pillars should be *translated* into domain-specific pillars, promoting the definition of trustworthiness metrics for each specific AI domain. Such a translation should also account for the current bias of EU trustworthiness pillars towards subsymbolic AI systems—where, for instance, the black-box nature of all components is given as understood when dealing with issues such as transparency, explainability, human interaction, even though it mostly concerns subsymbolic components only.

Thus, in the remainder of this paper we define the pillars of trustworthiness for AI systems based on Neuro-Symbolic (NeSy) integration. We analyse the seven EU-defined trustworthiness criteria for AI, and translate each of them into its NeSy counterpart, leveraging on the aspects of NeSy that promote fairness and explainability by design. On the other hand, leveraging both symbolic and subsymbolic paradigms, NeSy systems may be affected by robustness, safety, and bias issues from both sides – i.e., symbolic and subsymbolic –, hindering their overall trustworthiness. Therefore, a fundamental issue in the NeSy context is to identify whether and to what extent the blending of symbolic and subsymbolic techniques can either help or hinder trustworthiness, in particular in the perspective of the definition of ad-hoc trustworthiness metrics.

a) Human agency and oversight: in its original formulation this requirement stresses the need for the introduction of oversight mechanisms enabling informed interaction between AI agents and humans counterparts. The underlying assumption here is that humans can not understand AI system at all – or, can understand interaction with AI systems in a very limited way – so AI systems can never be considered as trustworthy, as human agents are incapable to fix the AI system when issues arise. When taking into account NeSy mechanisms, the symbolic and subsymbolic fusion component clearly affects the interaction with with its human counterpart. Instead, the symbolic component could represent the enabling agent for meaningful interaction between human and the system, promoting human-in-the-loop, human-on-the-loop, and human-in-command approaches.

NeSy version

The need for assessing to what extent the symbolic and subsymbolic *interaction* of NeSy components helps *improving* informed human-AI interaction and human oversight.

b) Technical robustness and safety: in its original formulation this requirement stresses the need for accuracy, reliability, resilience, and security of AI agents. Indeed, an inaccurate or unstable AI agent can not be considered trustworthy, as its behaviour may fluctuate radically throughout its life cycle. Let us consider for instance adversarial examples [10], [11], where slight perturbations of the input fed to the AI system result in radically different outcomes: AI system of that sort are inherently unreliable—thus untrustworthy. Even though this has motivated some research efforts focused on the identification of robustness issues of ML/DL systems, very small light has been shed on the robustness and safety issues of NeSy systems. NeSy agents rely on both symbolic and subsymbolic components, the former being – with some exception – verifiable and stable by design while the latter lacks of stability, verifiability or strong mathematical modeling of their behaviour and properties. The interaction of such elements introduce non-trivial behaviour in NeSy systems, where the symbolic components can be used as a helping tool for stabilising subsymbolic elements or the subsymbolic tools can be used to produce imperfect – thus unreliable – symbolic knowledge. Therefore, we consider relevant studying to what extent the verifiability of symbolic components alters during the integration process, and how the (in)stability of the subsymbolic element is impacted by the symbolic knowledge.

NeSy version

The need for assessing the impact of both symbolic (verifiable) and subsymbolic (not verifiable) *interaction* on the *stability* of the NeSy system.

c) Privacy and data governance: in its original formulation this requirement stresses the need for legitimate access to data, while taking into account data quality and integrity. This requirement identifies the untrustworthy nature of systems optimised over unreliable data, and promotes the introduction of open data for testing AI systems and their behaviour. To this end, NeSy systems differ quite heavily from their pure subsymbolic AI counterparts, as they – in most cases – require the processing of symbolic knowledge and data at the same time. Therefore, it is relevant to notice that data quality issues extend to knowledge quality issues when considering NeSy systems—even though symbolic knowledge is typically managed explicitly by AI programmers and is often verifiable in automatic way.

NeSy version

The need for ensuring the *quality* of both *data* and *symbolic knowledge* of a NeSy system, along with its accessibility.

d) Transparency: in its original formulation this requirement stresses the need for producing explanations of the AI agents' decision processes, deeming as untrustworthy those AI systems for which it is complex or unfeasible to obtain an explanation of its decision process. The definition of an AI system transparency depends on the complexity of the process of obtaining explanations, and their understandability. Indeed, in most AI scenarios multiple explanations can be drawn to render transparent the system at hand, depending on the level of detail needed and the process used. While being conceptually similar, the transparency level of NeSy systems – with respect to their pure subsymbolic AI counterparts – may differ a lot in terms of extraction complexity and understandability. Indeed, most NeSy systems represent a more transparent solution by design, as they leverage symbolic components, inputs or outputs, which are – to some extent – intrinsically understandable by humans. Therefore, we consider relevant to assess if – and to what extent – the integration components of NeSy systems impacts the transparency of the obtained agent(s).

NeSy version

The need for assessing the *gain* in terms of *transparency* obtained by a NeSy system with respect to its pure subsymbolic components.

e) Diversity, non-discrimination, and fairness: in its original formulation this requirement stresses the need for avoiding unfair bias and enable everyone's access to the AI technology. Indeed, biased AI technologies must not be deployed as they have been proven to increase the chance of harmful events against human agents. Given the relevance of fairness, several efforts have been put in place to investigate the nature of AI mechanisms' bias. However, biases of pure subsymbolic models and their NeSy counterparts differ conceptually in terms of their root causes: bias can rise in NeSy models as the consequence of any unexpected behaviour of their subsymbolic components, or their interaction with their symbolic elements. Indeed, similarly to what done for NeSy robustness, here it is relevant to highlight that the bias and fairness of symbolic components represent a verifiable and provable variable, while its interaction with subsymbolic elements does not, as it is not possible to define a-priori how the subsymbolic interaction impact the overall system behaviour. Therefore, it is fundamental for NeSy systems to consider possible biases rooted in each step of the fusion between symbolic and subsymbolic components. This is also valid for possible bias benefits that can be obtained from the interaction between symbolic and subsymbolic components in NeSy, as the symbolic elements can be used to tune the subsymbolic components to avoid biases that may arise during their optimisation.

NeSy version

The need for *measuring* biased and discriminative behaviour of NeSy agents rooted in the *interaction* between their symbolic and subsymbolic components.

f) Environmental and societal well-being: in its original formulation this requirement stresses the need for sustainability of AI agents and the transition to their environmentally friendly development, deeming as untrustworthy those AI agents that do not benefit all human beings, including future generations. While measuring the impact of pure subsymbolic AI agents on the environment has been the focus of several works in the AI community, the in-depth analysis of how NeSy mechanism can help reducing the environmental impact of AI. The symbolic component of several NeSy mechanism can be leveraged as a helping tool for reducing the amount of resources required for the optimisation of its subsymbolic component. Moreover, it is also possible for some NeSy mechanism to leverage symbolic approaches to achieve comparable performance – w.r.t. pure subsymbolic AI agent – while requiring a smaller memory footprint—resulting in smaller latency and energy consumption. On the other hand, the complex interaction between symbolic and subsymbolic components may introduce an overhead in the NeSy system, causing the waste of resources and thus decreasing the efficiency of the agent. Therefore, it is necessary to define a novel resource efficiency requirement for NeSy agents.

NeSy version

The need for assessing the *gain* in terms of *sustainability* of NeSy systems with respect to their pure subsymbolic components.

g) Accountability: in its original formulation this requirement stresses the need for mechanisms that ensure responsibility and accountability for AI systems and their outcomes. At its core, accountability can be defined as an obligation to inform about, and justify the AI's conduct [12]. Therefore, the fundamental property for AI's accountability is represented by *answerability*, which is the property of an AI system to allow for *interrogation* concerning a decision process. Accountability is closely tight to transparency, as it requires for an AI system to produce justification – a.k.a. explanations – for its actions. Therefore, a similar analysis to the one done for transparency applies to this context, where we stress the relevance of analysing the accountability gains obtained through symbolic and subsymbolic integration in NeSy systems over ML/DL counterparts.

NeSy version

The need for assessing the *gain* in terms of *answerability* obtained by a NeSy system with respect to its pure subsymbolic components.

III. On the Relevance of Trustworthiness Metrics

The trustworthy requirements proposed for both general AI systems and NeSy agents represent a general umbrella of concepts that should be covered in the system at hand. Indeed, none of the requirements defined so far give a specific characterisation of a target level – e.g., target fairness – for that requirement to be considered satisfied. Such general characterisation of trustworthiness is mainly caused by two contributing factors, namely

- *High variability characterising AI systems*. AI agents optimised to solve different tasks are expected to differ largely in terms of inner working principles. Therefore, identifying a common trustworthiness definition with the due level of detail represents a complex task
- *Conceptual complexity of trustworthiness building blocks*. Trustworthiness is defined as a collection of diverse features of a systems to be achieved for it to be worthy of humans' trust. However, some – if not most – of the trustworthiness sub-components are not easy-to-grasp concepts in their definition. For example, taking into account bias, we immediately understand that bias must be one of the sub-components required to achieve trustworthiness. However, the definition of bias by itself represents a complex task that have bogged researchers with troublesome questions like what is bias?, when is a system biased?, what is the minimum amount of bias for a system to be considered as such?. Being complex in their definition, these building blocks are also complex to measure effectively, hindering the overall level of trust measurement.

The issues connected with the general characterisation of AI trustworthiness hinder the applicability of such trustworthiness requirements. Indeed, while representing a valid starting point for analysing AI trustworthiness, these requirements do not fully allow to comprehensively grasp the extent of a system's trustworthiness. To this end, the definition of trustworthiness metrics – rather than requirements or pillars – represents an open issue of the utmost importance. Trustworthiness metrics make it possible to evaluate the extent of a system trust, allowing for a more detailed classification of the AI components to be deployed and the ones to block. However, the definition of a single general, flexible, and ubiquitous trustworthiness metric is made almost impossible by the same issues that affect the generality of trustworthiness requirements. Therefore, we here consider to translate the trustworthiness requirements into a set of equivalent trustworthiness metrics, taking into account the *high variability characterising AI systems* and the *conceptual complexity of trustworthiness building blocks*.

We first consider the issue connected with the *high variability characterising AI systems*. To enable the definition of rigorous trustworthy metrics, we here propose to consider the transition from the general AI trustworthy requirements to the corresponding pillars for each AI branch. Section II presents a similar transition from trustworthy AI into trustworthy NeSy. A similar transition can be identified for each and every AI domain, obtaining domain-specific detailed trustworthiness requirements. This step enables a stricter definition of trustworthiness for each AI domain, making it possible to focus more specifically on the peculiar approaches, components, and

aspects that characterise the domain under analysis.

To tackle the *conceptual complexity of trustworthiness building blocks*, we here propose to avoid focusing on the proposal of single, overly-complex trustworthy metrics with the aim of obtaining a general formulation applicable to any AI system. Rather, we suggest to tackle the measurement of systems' trustworthiness through the adoption of a broad set of highly-specialised metrics that analyse single components of the trustworthiness definition. In this context, we consider proposing a single metric or a set of metrics for each pillar/requirement of trustworthiness. The proposed metrics should focus on a specific issue or feature of the AI system at hand – such as its robustness to specific input perturbation, or the bias towards a specific group –, producing as output a single numeric value, describing its safety level—i.e., how much that issue is alarming for the system. Highly-specialised metrics can then be arbitrarily combined to obtain a dynamic trustworthiness score, depending on the trustworthiness components that are to be considered more relevant for the scenario under examination. This simplified process allows not just the easier definition of each set of trustworthiness metric – e.g., bias metrics, robustness metrics, etc. –, but also the evaluation of set based on a given relevance. Consider for example a scenario where the bias requirement should be considered as more relevant w.r.t. the human oversight requirement. Our approach allows a higher weight to be assigned to the bias metrics before its combination with the human oversight metrics to obtain the general trustworthy measurement. Therefore, we here propose to tackle the trustworthiness measurement issue by adopting a dynamic broad set of highly specific metrics that can be combined depending on the given measurement requirements.

IV. Background on SKI and SKE

In this section we provide an overview of the two NeSy mechanisms we focus on, namely Symbolic Knowledge Injection (Section IV-A) and Symbolic Knowledge Extraction (Section IV-B).

A. Symbolic Knowledge Injection (SKI)

Symbolic Knowledge Injection (SKI) defines the set of NeSy systems characterised by explicit procedures aiming at affecting how subsymbolic components draw their inferences for them to be made consistent with some given symbolic knowledge. In their definition, SKI mechanisms require having a subsymbolic predictor – a.k.a. model – and a given symbolic knowledge which always hold true for the considered context. The given symbolic knowledge should consist of logic formulæexpressed in any logic language of choice. In their scope, SKI mechanisms are designed to either *(i)* leverage the given input symbolic knowledge to enrich the training of the subsymbolic predictor; *(ii)* process the given symbolic knowledge via subsymbolic computations to achieve a novel, more meaningful symbolic knowledge; *(iii)* combine both of the previous processes. To achieve any of these scopes, SKI requires the given symbolic knowledge to be converted into a specific numeric form processable by the subsymbolic potion of the NeSy mechanism, to enable the injection process. More in detail, the converted symbolic knowledge is leveraged by the SKI approach to steer the learning process of the underlying subsymbolic model in any of the following way: *(i)* penalising the subsymbolic component during its training, whenever it violates the given symbolic knowledge, usually through defining a custom-made hybrid loss function; *(ii)* construct (a portion of) the subsymbolic component in such a way to make it reflect the given symbolic knowledge; *(iii)* convert the given symbolic knowledge into numeric-array form to be used as training data for the subsymbolic components of the NeSy system. In other words, SKI can be seen as the proces of optimising subsymbolic predictors in such a way that they are helped by the given symbolical knowledge.

B. Symbolic Knowledge Extraction (SKE)

Symbolic Knowledge Extraction (SKE) represents the set of NeSy approaches accepting subsymbolic predictors as input and producing symbolic knowledge as output. More in detail, SKE mechanisms aim at distilling the knowledge that a subsymbolic predictor has grasped from data into symbolic form, expressed by a set of logic formulæ. SKE enables the construction of a symbolic surrogate model that mimics the behaviour of a subsymbolic component. The obtained symbolic rules may then be exploited to either *(i)* understand and explain the behaviour of the original predictor; or *(ii)* replace subsymbolic components of the system while retaining its learning capabilities; To achieve symbolic knowledge construction, SKE can either *(i)* inspect (even partially) the parameters of the subsymbolic component – i.e., decompositional approaches –; or *(ii)* rely solely on the subsymbolic component's outputs—i.e., pedagogical approaches. Depending on the SKE approach the obtained symbolic knowledge can be under the form of lists of rules, decision trees or decision tables, each of them composed by any statement structure, such as propositional rules, fuzzy rules or any other kind of logic formulæ. In other words, SKE can be seen as the process of optimising symbolic AI components in such a way that their behaviour mimics given subsymbolic components.

V. NeSy Metrics for Trustworthiness

In this section we present the trustworthiness metrics (both available and missing ones) for NeSy systems, specifically focusing on SKI and SKE. We analyse each of the seven trustworthiness pillars/requirements separately to obtain a thorough representation of the state-of-the-art and future directions.

A. Human Oversight

NeSy version of human oversight requirement is defined as the need for assessing to what extent the symbolic and subsymbolic *interaction* of NeSy components helps *improving* informed human-AI interaction and human oversight.

1) Available Metrics: Most approaches to measure human oversight in AI scenarios focus on aspects of human-AI interaction, where explanation of behaviours represents the most important component of the interaction process. As a

results, much attention has been paid to the measurement of how explanations could guide people to respond to and predict the AI system behaviour [13]. A large number of studies exist in this realm, which mainly leverage on users to subjectively rate system predictability, likability, etc.[14] While useful in order to define systems predictability, these studies lack the assessment of human influence and control on the AI system at hand. The reason for this is to be found mainly on the black-box and data-driven nature of subsymbolic models that these works take into account. Indeed, most – if not all – subsymbolic models allow for limited control by the human users, given mostly by the data gathering and selection process.

2) Missing Metrics: Unlike pure subsymbolic systems, NeSy models intrinsically enable higher level of human oversight via the integration of symbolic knowledge. However, the extent of such oversight capabilities should be studied in depth through the proposal of ad-hoc metrics that measure how much the behaviour of a NeSy system can be controlled by a human user. To this aim, in the SKI context, we consider proposing a novel metric assessing the impact of the injection process to the underlying model. The impact can be measured as the amount of injected knowledge that is effectively absorbed by the underlying model. The metric would assess the level of available human oversight in SKI systems, allowing for a precise definition of the extent of human control. Meanwhile, the SKE context emphasises the need for measuring the modifiability of the extracted symbolic knowledge from an initial subsymbolic predictor. Indeed, SKE approaches by themselves do not allow for an in-depth control of the model behaviour, but rather enable their inspection. In this context, a desirable solution is represented by refining the extracted knowledge and using it as input for a SKI system acting upon the same subsymbolic model. This process would enable a sort of debugging loop of NeSy systems leveraging both SKE and SKI, with an increased potential for human oversight. Here, we require the definition of an ad-hoc metric capable of assessing the portion of symbolic knowledge that can be extracted, refined and injected back in the system with it being correctly assimilated by the model.

B. Robustness

NeSy version of the robustness requirement is defined as the need for assessing the impact of symbolic (verifiable) and subsymbolic (not verifiable) *interaction* on the *stability* of the NeSy system.

1) Available Metrics: The state-of-the-art picture of NeSy robustness emphasises the lack of a common agreement on the definition of robustness itself, thus leading to diverging works focusing on opposite aspects of NeSy systems. Indeed, in this context, several works focus on highlighting the robustness of NeSy models in terms of their performance over complex or out-of-distribution inputs [15], [16], [17]. Although relevant for pointing out the potential of NeSy approaches, these works propose somehow misleading definitions of robustness, mostly focusing on NeSy flexibility rather than its stability. NeSy systems may perform well on complex and out-of-distribution samples, while suffering instability on small input perturbations—causing robustness collapse. Several other concepts have been taken into account when considering NeSy robustness such as prediction coherence and consistency [18], subsymbolic verification through neuro-symbolic integration [19], avoidance of reasoning shortcuts [20] and many more. However, the majority of these approaches not only assess an ad-hoc concept of robustness, but also focus on its qualitative evaluation thus failing to assess the quantitative aspect required to achieve robustness metrics.

While it is true that there exists some confusion concerning the definition of NeSy robustness, there are few relevant works aiming at defining precise robustness metrics. More in detail, Yang et al. [21] present a novel learning approach for neuro-symbolic programs, showing its robustness against input perturbations in terms of provably safe portion of the learned model. In this context, NeSy robustness against adversarial attacks represents a popular area of research with several works aiming at proving either qualitatively [22] or quantitatively [23] the safety of NeSy approaches. Most of these works define robustness in terms of accuracy degradation over varying input perturbation intensity, independently of the input perturbation type and magnitude.

2) Missing Metrics: As a result of the mixed focus given to NeSy aspects when tackling robustness, several aspect of NeSy robustness and stability have not been thoroughly analysed, yet. Indeed, there exists the need to study if – and to what extent – the stability and verifiability of symbolic AI components is preserved throughout the integration process in NeSy models. In this context, focusing on the SKI realm, we suggest that a measure of integration stability – as the portion of symbolic elements that are correctly integrated in the injected model – is needed here. Such a metric would basically represent the portion of symbolic control that a NeSy system can attain during its integration step. Secondly, also those scenarios where the symbolic elements of NeSy models suffer from some sort of imperfection have to be taken into account. Here, it is important to measure the stability of SKI models when the injected knowledge is altered as a result of some imperfect automation process. Finally, it is also relevant to measure the stability of NeSy systems over symbolic representation variability, to assess how different symbolic representations – e.g., logic formulæ, knowledge graphs, etc. – may impact the integration process. To this end, we propose to measure the performance of SKI integration when two syntactically different yet equivalent chunks of symbolic knowledge are exploited in the same integration process.

C. Data & Knowledge Quality

NeSy version of the data & knowledge quality requirement is defined as the need for ensuring the *quality* of both *data* and *symbolic knowledge* of a NeSy system, along with its accessibility.

1) Available Metrics: Given the impact of data quality on the optimisation process of ML and DL systems, several

quality metrics are available, namely: *(i)* class overlap [24], *(ii)* boundary complexity [25], *(iii)* label noise [26], *(iv)* class imbalance [27], *(v)* missing value analysis [28], and many more. Although designed for subsymbolic AI models, these metrics translate to the data-driven component of NeSy systems without particular issues, especially in those systems that follow a neural to symbolic – neuro $\rightarrow$ symbolic [29] – pipeline such as SKE approaches. In this context, these metrics makes it possible to check the correctness of the information that the subsymbolic components of NeSy gather from the data.

2) Missing Metrics: Unlike pure subsymbolic approaches – which rely solely on data for optimisation –, NeSy models gather information from both a data-driven and a symbolic knowledge component. In this context, it is fundamental to assess the level of compatibility or overlap between the data and the symbolic knowledge to be combined. In most NeSy systems quite a strong overlap is required between data and symbolic knowledge in order to avoid optimisation drift issues, where the integrated knowledge contrasts concepts learnt from the data. Meanwhile, a perfect overlap would also not be ideal in NeSy systems, as the optimisation process would gather the same information from both data and symbolic knowledge. Therefore, we here stress the need for new metrics that could measure the conceptual and technical overlap between data and symbolic knowledge at hand. Another relevant aspect to measure in this context is represented by the quality of the symbolic component of the NeSy system. While symbolic AI approaches are verifiable and deemed trustworthy, several NeSy – especially SKI – approaches rely on the integration of knowledge bases given a-priori and defined by human experts. Although mostly reliable, knowledge bases may be either incomplete or imperfect due to the human-centred building process. Therefore, metrics are needed that would make it possible to score knowledge components exploited in NeSy systems.

D. Transparency

NeSy version of the transparency requirement is defined as the *transparency gain* obtained by a NeSy system with respect to its pure subsymbolic components.

1) Available Metrics: When focusing on transparency, most of the available metrics for AI and NeSy models focus on explanations quality evaluation. Generally speaking, explanations quality is characterised by several key attributes [30], namely: *(i)* understandability – i.e., explanation complexity –; *(ii)* completeness – i.e., explanation coverage –; *(iii)* sufficiency of detail – i.e., explanations depth –; *(iv)* usefulness – i.e., explanation applicability –; and *(v)* feeling of satisfaction—i.e., explanation interactivity. By focusing on some of the above attributes, several works propose explainability and transparency metrics for AI and NeSy. Authors in [31] introduce a set of metrics to evaluate interpretability methods through measurements of simplicity, broadness, and fidelity of explanations. Meanwhile, Holzinger et al. [32] introduce a system causability scale to measure explanations quality, based on the notion of causability [33] together with the notion of usability scale. Although designed for explanations in general, these metrics nicely fit in the SKE frame, where they can be used to assess the quality of the extraction mechanism, as done in [34], where authors focus on unambiguity, interpretability, and interactivity of explanations.

2) Missing Metrics: Available explainability metrics aim at measuring the quality of explanations in absolute terms—i.e., how good are my extracted explanations? Meanwhile, our definition of NeSy transparency requires to measure the *gain* in transparency obtained from symbolic and subsymbolic integration. Therefore, there is the need for novel metrics for NeSy systems comparing the quality of a system's explanations before and after symbolic and subsymbolic integration. Moreover, we here stress the unbalanced nature of explainability metrics, as most metrics focus solely on features of explanations that are automatically measurable – e.g., correctness, coverage, length, etc. –, whereas there are basically no metrics focusing on human oriented specifications. A relevant issue for future research in this are is the definition of metrics that account for the subjective human factor in explanations, assessing the level of explanations satisfaction and understandability via human-assisted experimentation. Finally, it should be noted that transparency should not just focus on measuring the quality of the explanations that can be obtained from a system, but should instead assess the complexity of the process for extracting those explanations, too. Indeed, explanations obtained from a DL model using SKE may be complete, understandable and useful, but require a high computational burden to be extracted, rendering the overall DL and SKE process less transparent.

E. Fairness

NeSy version of fairness requirement is defined as the need for *measuring* biased and discriminative behaviour of NeSy agents rooted in the *interaction* between their symbolic and subsymbolic components.

1) Available Metrics: Given the nuances characterising a context-dependent notion like fairness, developing quantitative formulations for fairness metrics is challenging [35]. In the general context of AI systems, fairness is generally regarded as *outcome fairness*, which is the definition of equality of the decision making process outcomes. Here, fairness can be categorised into individual vs. group notions of fairness, and observational vs. causal approaches to assess fairness [36]. Observational fairness approaches are characterised by a number of existing metrics, such as: *(i) independence metrics* – e.g., statistical parity, group fairness, demographic parity, etc. –; *(ii) separation metrics* – e.g., equal opportunity, equalised odds, predictive equality, etc. –; and *(iii) sufficiency metrics*—e.g., groups calibration, predictive parity, etc.

While representing a fundamental requirement, fairness in NeSy setups is yet to be explored in detail. Indeed, only a handful of works have investigated fairness in NeSy systems. Authors in [37] propose to leverage the combination of symbolic knowledge extraction from Logic Tensor Networks [38]

and injection of fairness constraints via continual learning to enforce fairness. Gao et al. [39] inject a fairness-based component in the loss function of subsymbolic models during their optimisation process to achieve higher fairness. Beyond their obvious relevance, these work focus solely on possible fairness benefits obtained through NeSy, as they rely on the application of SKI and SKE to reduce bias issues, leveraging the general AI fairness metrics. Therefore, available NeSy-specific fairness metrics are still missing that would aim at measuring just the impact of symbolic and subsymbolic integration upon fairness. This deficit is probably due to two aspects: *(i)* most observational fairness metrics are considered to be applicable to NeSy systems without modification; and *(ii)* most research focuses on measuring the fairness and assess it, rather than aiming at identifying its root causes.

2) Missing Metrics: In its NeSy version, the fairness requirement highlights the need to assess the possible fairness issues or improvements that arise from the use of symbolic and subsymbolic integration. It is clear that this requirement is not satisfied by available fairness metrics. Indeed, although most observational fairness metrics apply to NeSy systems, they do not allow for identification of the root causes of bias. One approach to tackle this issue would be to measure NeSy fairness as a differential of observational fairness between a SKI/SKE model and its ML/DL counterpart. However, such an approach would be over-simplistic, as it would not allow the specific sub-components of the integration process or of the symbolic knowledge that impact fairness to be captured. One possible solution would be to measure the fairness of NeSy systems over a set of symbolic knowledge bases, each representing a specific set of fairness goal. This process would allow fairness goal to be decomposed into its components/elements, then measure how well a NeSy system can enforce each fairness element.

F. Resource Efficiency

NeSy version of the resource efficiency requirement is defined as the need for assessing the *gain* in terms of *sustainability* with respect to pure subsymbolic counterparts.

1) Available Metrics: When dealing with resource efficiency of AI systems in general, the detailed definition of the set of resources to take into account represents a fundamental aspect. Several elements of the system at hand can be identified as resources, ranging from the energy required by the system to be optimised to its scalability—e.g., overall complexity. In this context, Agiollo et al. [40], [41] propose a rigorous definition of resource efficiency improvements achievable by SKI systems spanning over four different resource components. More in detail, the authors focus on the definition of energy, latency, memory, and data efficiency of any SKI model, aiming at addressing its environmental impact – e.g., energy and data –, and its scalability—e.g., memory and latency. These metrics are defined as the relative difference in terms of resources – e.g., energy, etc. – required to optimise a SKI model to reach the same level of performance of a subsymbolic counterpart. To this end, the authors define each of the resource analysed, and provide for a tool to measure them in a SKI setup, showing how SKI can improve energy and data efficiency, while degrading the system latency. Latency increments are linked with the increased complexity of the system given by the interaction between symbolic and subsymbolic components, which is however beneficial in terms of number of data required for optimising the model. Indeed, several other works show the data efficiency of NeSy models – such as [42], [43], [44] – even though they lack a proper definition for efficiency.

2) Missing Metrics: As data efficiency represents one of the declared advantages of NeSy systems, most of the literature focuses specifically on this aspect, leaving some space for investigation about other relevant aspects of resource efficiency. More in detail, detailed analysis of the environmental impact of AI and NeSy models development in terms of their carbon footprint are still mostly missings. Studying the energy consumption of the development of a single NeSy model is not enough, as the computation infrastructure used throughout this development – such as clusters and cloud infrastructures – strongly impact its environmental footprint. Moreover, whereas few metrics exist that assess the efficiency of NeSy under the SKI perspective, there are basically no metrics for resource efficiency in the SKE area. In this context, it would be desirable to have metrics similar to the ones obtained for SKI comparing the resource usage of the original subsymbolic model and its symbolic emulation. Depending on the SKE approach at hand, it is possible to consider extracting a small symbolic AI models mimicking the behaviour big DL frameworks. The small symbolic model obtained may help hugely reducing the amount of resources – especially energy, latency, and memory – required to deploy the AI system. Therefore, we here suggest as a future direction to investigate whether – and to what extent – SKE can produce small and fast counterparts of DL models. Here, the resource efficiency metric could be simply designed as the relative difference between the amount of resources required to run the original DL model and its symbolic emulation.

G. Accountability

NeSy version of the accountability requirement is defined as the need for assessing the *gain* in terms of *answerability* obtained by a NeSy system with respect to its pure subsymbolic components.

1) Available Metrics: As it is represented by the answerability of an AI system, accountability is closely tight to transparency. Indeed, accountability requires the underlying system to be explainable, and the explanations to be correct, reliable, and comprehensible. Correctness and reliability of explanations depend on the precision of the AI system and its explanation construction counterpart. Therefore, most efforts in this field focus on the explainability of the AI/NeSy system at hand. As a result, the set of available AI and NeSy metrics for accountability is basically represented by the same set of metrics presented in Section V-D.

2) Missing Metrics: While being tightly linked with explainability, accountability also requires the extracted explanations to be correct and reliable. As correctness and reliability mostly depend on the precision of the AI/NeSy system, we here propose to define novel accountability metrics by opportunistically mixing transparency metrics (Section V-D) and robustness metrics (Section V-B). Therefore, accountability metrics should be defined as the result of explainability metrics applied over a set of input perturbations, measuring the rate of change of the obtained explanations.

VI. Conclusions

Trustworthiness of AI systems represents a fundamental requirement for their ubiquitous deployment. The notion of Trustworthy AI as defined by the EU is mostly a general one, yet implicitly accounting for issues coming from popular ML and DL techniques—so it fits well subsymbolic AI systems. A set of novel NeSy systems calls for a more specific definition of trustworthiness, as they rely on the integration of subsymbolic and symbolic AI where the symbolic components may affect – either positively or negatively – the trust level of the system. Accordingly, in this paper we analyse how the AI trustworthiness requirements defined by the EU translate to the NeSy realm, focusing on the relevant elements of the NeSy integration process impacting trust. First we analyse in detail each pillar of trust and its implication on NeSy models, then we focus on the available metrics for measuring such requirements. The state-of-the-art analysis highlights a lack of available metrics for most trustworthiness aspects when specifically considering NeSy systems. Therefore, we suggest potential future directions to explore in the analysis of NeSy trust along with with related metrics definitions. We believe that the rigorous definition of novel trust metrics tailored to NeSy systems is going to represent an essential step towards measurably reliable and trustworthy AI systems based on neuro-symbolic integration.

Acknowledgment

This work was partially supported by PNRR – M4C2 – Investimento 1.3, Partenariato Esteso PE00000013 – "FAIR—Future Artificial Intelligence Research" – Spoke 8 "Pervasive AI", funded by the European Commission under the NextGenerationEU programme, and by the CHIST-ERA IV project "Expectation" – CHIST-ERA-19-XAI-005 –, co-funded by EU and the Italian MUR (Ministry for University and Research).

References

[1] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8627998

[2] A. Agiollo, G. Ciatto, and A. Omicini, "*Shallow2Deep*: Restraining neural networks opacity through neural architecture search," in *Explainable and Transparent AI and Multi-Agent Systems*, ser. Lecture Notes in Computer Science, D. Calvaresi, A. Najjar, M. Winikoff, and K. Främling, Eds. Cham: Springer, 2021, vol. 12688, pp. 63–82. [Online]. Available: http://link.springer.com/10.1007/978-3-030-82017-6_5

[3] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 604–624, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9075398

[4] A. Agiollo, L. C. Siebert, P. K. Murukannaiah, and A. Omicini, "The quarrel of local post-hoc explainers for moral values classification in natural language processing," in *Explainable and Transparent AI and Multi-Agent Systems*, ser. Lecture Notes in Computer Science, D. Calvaresi, A. Najjar, A. Omicini, R. Aydoğan, R. Carli, G. Ciatto, Y. Mualla, and K. Främling, Eds. Springer, 2023, vol. 14127, ch. 6. [Online]. Available: http://link.springer.com/10.1007/978-3-031-40878-6_6

[5] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 249–270, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9039675

[6] A. Agiollo and A. Omicini, "GNN2GNN: Graph neural networks to generate neural networks," in *Uncertainty in Artificial Intelligence*, ser. Proceedings of Machine Learning Research, J. Cussens and K. Zhang, Eds., vol. 180. ML Research Press, Aug. 2022. ISSN 2640-3498 pp. 32–42, proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands. [Online]. Available: https://proceedings.mlr.press/v180/agiollo22a.html

[7] A. Agiollo, E. Bardhi, M. Conti, R. Lazzeretti, E. Losiouk, and A. Omicini, "GNN4IFA: Interest flooding attack detection with graph neural networks," in *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, IEEE Computer Society. Los Alamitos, CA, USA: IEEE Computer Society, Jul. 2023. ISBN 978-1-6654-6512-0 pp. 615–630. [Online]. Available: https://www.computer.org/csdl/proceedings-article/eurosp/2023/651200a615

[8] K. Benidis, S. S. Rangapuram, V. Flunkert, Y. Wang, D. C. Maddix, A. C. Türkmen, J. Gasthaus, M. Bohlke-Schneider, D. Salinas, L. Stella, F. Aubet, L. Callot, and T. Januschowski, "Deep learning for time series forecasting: Tutorial and literature survey," *ACM Computing Surveys*, vol. 55, no. 6, pp. 121:1–121:36, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3533382

[9] A. Agiollo, M. Conti, P. Kaliyar, T. Lin, and L. Pajola, "DETONAR: Detection of routing attacks in RPL-based IoT," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1178 – 1190, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9415869

[10] J. Zhang and C. Li, "Adversarial examples: Opportunities and challenges," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2578–2593, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8842604

[11] A. C. Serban, E. Poll, and J. Visser, "Adversarial examples on object recognition: A comprehensive survey," *ACM Computing Surveys*, vol. 53, no. 3, pp. 66:1–66:38, 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3398394

[12] C. Novelli, M. Taddeo, and L. Floridi, "Accountability in artificial intelligence: what it is and how it works," *AI & SOCIETY*, pp. 1–12, 2023. [Online]. Available: https://link.springer.com/10.1007/s00146-023-01635-y

[13] M. M. A. de Graaf and B. F. Malle, "How people explain action (and autonomous intelligent systems should too)," in *2017 AAAI Fall Symposia, Arlington, Virginia, USA, November 9-11, 2017*. AAAI Press, 2017, pp. 19–26. [Online]. Available: https://aaai.org/ocs/index.php/FSS/FSS17/paper/view/16009

[14] C. Huang and B. Mutlu, "Robot behavior toolkit: generating effective social behaviors for robots," in *International Conference on Human-Robot Interaction, HRI'12, Boston, MA, USA - March 05 - 08, 2012*, H. A. Yanco, A. Steinfeld, V. Evers, and O. C. Jenkins, Eds. ACM, 2012, pp. 25–32. [Online]. Available: https://dl.acm.org/doi/10.1145/2157689.2157694

[15] Z. Li, X. Wang, E. Stengel-Eskin, A. Kortylewski, W. Ma, B. V. Durme, and A. L. Yuille, "Super-CLEVR: A virtual benchmark to diagnose domain robustness in visual reasoning," *CoRR*, vol. abs/2212.00259, 2022. [Online]. Available: https://arxiv.org/abs/2212.00259

[16] C. W. Wu, A. C. Wu, and J. Strom, "DeepTune: Robust global optimization of electronic circuit design via neuro-symbolic optimization," in *IEEE International Symposium on Circuits and Systems, ISCAS 2021, Daegu, South Korea, May 22-28, 2021*. IEEE, 2021, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9401488

[17] A. Liu, H. Xu, G. Van den Broeck, and Y. Liang, "Out-of-distribution generalization by neural-symbolic joint training," in *AAAI Conference on Artificial Intelligence*, vol. 37, no. 10, 2023, pp. 12 252–12 259. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26444

[18] M. I. Nye, M. H. Tessler, J. B. Tenenbaum, and B. M. Lake, "Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning," in *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 25 192–25 204. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/d3e2e8f631bd9336ed25b8162aef8782-Abstract.html

[19] X. Xie, K. Kersting, and D. Neider, "Neuro-symbolic verification of deep neural networks," in *Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*, L. De Raedt, Ed. ijcai.org, 2022, pp. 3622–3628. [Online]. Available: https://www.ijcai.org/proceedings/2022/503

[20] E. Marconato, G. Bontempo, E. Ficarra, S. Calderara, A. Passerini, and S. Teso, "Neuro symbolic continual learning: Knowledge, reasoning shortcuts and concept rehearsal," *CoRR*, vol. abs/2302.01242, 2023. [Online]. Available: https://arxiv.org/abs/2302.01242

[21] 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. [Online]. Available: https://openreview.net/forum?id=NYBmJN4MyZ

[22] M. R. Vilamala, T. Xing, H. Taylor, L. Garcia, M. Srivastava, L. M. Kaplan, A. D. Preece, A. Kimmig, and F. Cerutti, "DeepProbCEP: A neuro-symbolic approach for complex event processing in adversarial settings," *Expert Systems with Applications*, vol. 215, pp. 119 376:1–26, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422023946

[23] G. Ibarra-Vázquez, G. Olague, M. Chan-Ley, C. Puente, and C. Soubervielle-Montalvo, "Brain programming is immune to adversarial attacks: Towards accurate and robust image classification using symbolic learning," *Swarm and Evolutionary Computation*, vol. 71, p. 101059, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210650222000311

[24] M. Denil and T. P. Trappenberg, "Overlap versus imbalance," in *Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science, A. Farzindar and V. Keselj, Eds., vol. 6085. Springer, 2010, pp. 220–231. [Online]. Available: https://link.springer.com/10.1007/978-3-642-13059-5_22

[25] A. C. Lorena, L. P. F. Garcia, J. Lehmann, M. C. P. de Souto, and T. K. Ho, "How complex is your classification problem?: A survey on measuring classification complexity," *ACM Computing Surveys*, vol. 52, no. 5, pp. 107:1–107:34, 2019. [Online]. Available: https://dl.acm.org/doi/10.1145/3347711

[26] C. G. Northcutt, L. Jiang, and I. L. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *Journal of Artificial Intelligence Research*, vol. 70, pp. 1373–1411, 2021. [Online]. Available: https://jair.org/index.php/jair/article/view/12125

[27] Y. Lu, Y. Cheung, and Y. Y. Tang, "Bayes imbalance impact index: A measure of class imbalanced data set for classification problem," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3525–3539, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8890005

[28] D. C. Corrales, J. C. Corrales, and A. Ledezma, "How to address the data quality issues in regression models: A guided process for data cleaning," *Symmetry*, vol. 10, no. 4, p. 99, 2018. [Online]. Available: https://www.mdpi.com/2073-8994/10/4/99

[29] M. K. Sarker, L. Zhou, A. Eberhart, and P. Hitzler, "Neuro-symbolic artificial intelligence," *AI Communications*, vol. 34, no. 3, pp. 197–209, 2021. [Online]. Available: https://content.iospress.com/articles/ai-communications/aic210084

[30] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: challenges and prospects," *CoRR*, vol. abs/1812.04608, 2018. [Online]. Available: http://arxiv.org/abs/1812.04608

[31] A. Nguyen and M. R. Martínez, "On quantitative aspects of model interpretability," *CoRR*, vol. abs/2007.07584, 2020. [Online]. Available: https://arxiv.org/abs/2007.07584

[32] A. Holzinger, A. M. Carrington, and H. Müller, "Measuring the quality of explanations: The system causability scale (SCS)," *KI - Künstliche Intelligenz*, vol. 34, no. 2, pp. 193–198, 2020. [Online]. Available: https://link.springer.com/10.1007/s13218-020-00636-z

[33] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, pp. e1312:1–13, 2019. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1312

[34] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Interpretable & explorable approximations of black box models," *CoRR*, vol. abs/1707.01154, 2017. [Online]. Available: http://arxiv.org/abs/1707.01154

[35] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii, "Matroids, matchings, and fairness," in *22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 2019, pp. 2212–2220. [Online]. Available: http://proceedings.mlr.press/v89/chierichetti19a.html

[36] R. Calegari, G. G. Castañé, M. Milano, and B. O'Sullivan, "Assessing and enforcing fairness in the AI lifecycle," in *32nd International Joint Conference on Artificial Intelligence (IJCAI 2023)*. Macau, China: IJCAI, August 19–25 2023.

[37] B. Wagner and A. d'Avila Garcez, "Neural-symbolic integration for fairness in AI," in *AAAI-MAKE 2021 – Combining Machine Learning and Knowledge Engineering*, ser. CEUR Workshop Proceedings, A. Martin, K. Hinkelmann, H. Fill, A. Gerber, D. Lenat, R. Stolle, and F. van Harmelen, Eds., vol. 2846. CEUR-WS.org, 2021. [Online]. Available: https://ceur-ws.org/Vol-2846/paper5.pdf

[38] S. Badreddine, A. S. d'Avila Garcez, L. Serafini, and M. Spranger, "Logic tensor networks," *Artificial Intelligence*, vol. 303, pp. 103 649:1–39, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370221002009

[39] X. Gao, J. Zhai, S. Ma, C. Shen, Y. Chen, and Q. Wang, "FairNeuron: improving deep neural network fairness with adversary games on selective neurons," in *44th International Conference on Software Engineering, ICSE 2022*. ACM, 2022, pp. 921–933. [Online]. Available: https://dl.acm.org/doi/10.1145/3510003.3510087

[40] A. Agiollo, A. Rafanelli, and A. Omicini, "Towards quality-of-service metrics for symbolic knowledge injection," in *WOA 2022 – 23rd Workshop "From Objects to Agents"*, ser. CEUR Workshop Proceedings, A. Ferrando and V. Mascardi, Eds., vol. 3261. Sun SITE Central Europe, RWTH Aachen University, 2022. ISSN 1613-0073 pp. 30–47. [Online]. Available: http://ceur-ws.org/Vol-3261/paper3.pdf

[41] A. Agiollo, A. Rafanelli, M. Magnini, G. Ciatto, and A. Omicini, "Symbolic knowledge injection meets intelligent agents: QoS metrics and experiments," *Autonomous Agents and Multi-Agent Systems*, vol. 37, no. 2, pp. 27:1–27:30, Jun. 2023. [Online]. Available: https://link.springer.com/10.1007/s10458-023-09609-6

[42] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=rJgMlhRctm

[43] Q. Zhang, L. Wang, S. Yu, S. Wang, Y. Wang, J. Jiang, and E. Lim, "NOAHQA: Numerical reasoning with interpretable graph question answering dataset," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. ACL, 2021, pp. 4147–4161. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.350/

[44] B. Škrlj, M. Martinc, N. Lavrač, and S. Pollak, "autoBOT: evolving neuro-symbolic representations for explainable low resource text classification," *Machine Learning*, vol. 110, no. 5, pp. 989–1028, 2021. [Online]. Available: https://link.springer.com/article/10.1007/s10994-021-05968-x