Neurosymbolic Artificial Intelligence 0 (2024) 1–0
IOS Press

A Neurosymbolic Approach to AI Alignment

Benedikt J. Wagner [a,*] and Artur d'Avlia Garcez [a]
[a] *Department of Computer Science, City, University of London, London, United Kingdom*
E-mail: Benedikt.Wagner@city.ac.uk

Abstract. We propose neurosymbolic integration as an approach for AI alignment via concept-based model explanation. The aim is to offer AI systems the ability to learn from human revision but also assist humans at evaluating AI capabilities. The proposed method allows users and domain experts to learn about the data-driven decision making process of large neural network models and to impose a particular behaviour onto such models. The models are queried using a symbolic logic language that acts as a lingua franca between humans and model representations. Interaction with the user then confirms or rejects a revision of the model using logical constraints that can be distilled back into the neural network. We illustrate the approach using the Logic Tensor Network framework alongside Concept Activation Vectors and apply it to Convolutional Neural Networks and the task of achieving quantitative fairness. Our results illustrate how the use of a logical language is able to provide users with a formalisation of the model's decision making whilst allowing users to steer the model towards a given alignment constraint.

Keywords: Neurosymbolic Alignment, Concept Representation, Knowledge Revision, Explainable AI

1. Introduction

As AI systems become increasingly relevant and popular in various domains, there is a growing concern about their potential misalignment with human values and goals. To address these concerns, researchers have advocated the development of transparent and interpretable AI models, capable of providing meaningful insights to users about the model's internal processes. Users, in turn, in particular system developers may also wish to be able to influence or control the system's decision-making processes. In this paper, we argue that the fusion of symbolic logic and neural network methodologies, termed neurosymbolic AI, is not only a theoretical advancement but a practical necessity for aligning AI systems with human ethical standards [1]. Specifically, we defend the position that formal logic, when integrated with neural networks, serve as a vital interface for both guiding AI learning processes and providing comprehensible explanations of AI decisions to users. This neurosymbolic approach, we propose, is essential in achieving a balance between the technical efficacy of AI systems and their alignment with human values, particularly in aspects of fairness and explainability. We argue that the neurosymbolic approach offers a pathway to the development of AI systems that are transparent and aligned with human goals. We will illustrate how symbolic logic can be used to query a deep neural network thus closing the neurosymbolic cycle as shown in Figure 1. We will use a practical application in computer vision and quantitative fairness. The use of a formal language as the interface between humans and AI, by contrast to natural language, will be illustrated as the way to offer an objective evaluation of results.

It has become clear that the current standard approach to alignment of AI systems - Reinforcement Learning with Human Feedback (RLHF) [1, 2] - is itself unethical and ineffective [3]. This is not just because of the exploitation

*Corresponding author. E-mail: Benedikt.Wagner@city.ac.uk.
[1] In this manuscript, the terms *Artificial Intelligence* (AI) and *Machine Learning* (ML) are used interchangeably. This reflects the current dominant role of machine learning within the AI field, particularly in AI alignment contexts.

of data labelers in poor countries but because of the much less mentioned and untested mental health impact on workers exposed to the worst examples of misalignment - hate speech, child sex abuse, etc. - for many hours a day. RLHF at scale is unsustainable.

A different approach is needed. Instead of alignment as a afterthought following large-scale training on massive, uncurated data, alignment should be a first-class concern as part of an incremental training process. This is promoted by the neurosymbolic cycle with training happening in stages, allowing knowledge to be validated as training progresses and consolidated back into the system.

The remainder of the paper is organised as follows. The next section lays the foundational concepts and assumptions. We then examine logic's role in bridging human-AI communication. In the subsequent section, we delve into aligning AI with human understanding. The *Logic Tensor Network (LTN) Framework* section illustrates LTN's application to AI alignment, and the following *Ethical Alignment with Fairness Constraints* section illustrates how we incorporate fairness in AI decisions. We then outline a practical neurosymbolic approach to concept alignment in computer vision, and the conclusion summarises our findings and suggests directions for future research.

2. Background and Related Work

At the core of our approach is the assumption that concepts described as logic predicates are fundamental to both the structure and functionality of neurosymbolic AI systems. Logic predicates serve as the building block for representing complex ideas and relationships within an AI system's reasoning process. Concepts, extracted from data and represented symbolically, allow humans to interact and intervene in the learning system, offering a measurable approach to alignment between AI and human understanding. This assumption underpins our belief that a neurosymbolic AI system, to be aligned with human values, must be capable of reasoning with and about symbolic logic predicates and concepts.

We assume that symbolic representation, as opposed to purely sub-symbolic approaches, is crucial for creating AI systems that are both explainable and aligned with human values. Symbolic representation allows for a more transparent and interpretable decision-making process, enabling users to understand and trust an AI system's outputs. The relationships between external entities and representations and symbols are bidirectional, meaning that external entities affect representation and symbols, but symbols also affect how we act in the world [4].

In the area of knowledge extraction, explainable AI and mechanistic interpretability, there is a tension between the desire for logical precision and practical necessity. While precision may be desirable, it is often impractical from a computational complexity perspective in the context of the symbol grounding problem. Within the domain of computer vision, for example, this challenge becomes especially salient. Explanations rooted in logical relations of pixel values may offer precision, but risk being too complicated, at the wrong level of abstraction, and therefore useless in practice. Our aim is to harness logic at higher levels of abstraction, such as objects, shapes and colours, thereby mirroring human explanatory practices. Additionally, we seek to imbue the model with behaviour that operates at this level of abstraction, yielding a more intuitive experience for users.

Another fundamental assumption is the necessity for AI systems to be interactive and adaptable. This means that AI systems should not only learn from data but also from interactions with users. Such interactions, often provided in the form of feedback to the AI system's reasoning and outputs, should allow the system to continually refine, aligning its operation with human expectation and ethical standards.

Futia et. al [5] underline the importance of neurosymbolic integration for explainability by suggesting that traditional explainable AI (XAI) methods lack the ability to provide explanations for the variety of target audiences. While most of the explainability methods may be valuable at providing insight to Machine Learning (ML) experts, domain experts in application areas such as finance or healthcare may struggle to interpret most forms of explanation. It is proposed that *interactive integration with semantically-rich representations is key to refining explanations targeted at different stakeholders* [5]. Indeed, having flexibility in the abstract representation of information that forms an explanation is key to leveraging domain expertise through interaction and revision of the decision making process. More specifically, in this case, neurosymbolic integration should allow us to overcome the static nature of the current ML paradigm. With very few exceptions, e.g. [6], explainability methods of today do not provide an ability to act on extracted information. Upon finding an undesired property of the system, in general the only way

to influence the model is to retrain it until a better model is found. However, retraining as an unguided process can only be influenced indirectly by an extracted explanation, normally through the collection of additional data. The result is that many XAI methods may become limited in their usefulness, with retraining commonly resulting in catastrophic forgetting of previously acquired information.

By closing the neurosymbolic cycle, the objective is to guide further network training informed by extracted explanations and, in the process, foster a deeper alignment between system and human expectations and values, focusing on fairness in the case of this paper. This integration seems to be not only a theoretical advancement but also a practical necessity for aligning AI with human ethical standards through the provision of new methods that allow users and domain experts to understand and influence the decision-making process of large neural networks. In [7], a model is introduced that maps inputs to concepts and then to targets for interpretability and intervention. In [8], emphasis is placed on user interactions with the learning process to enhance model transparency and trust. By comparison, we integrate user interaction within the AI alignment process. The goal is to ensure that the AI system is transparent but also adaptable to user feedback.

The recent focus of ML towards concept-based models [9–13] along with the emergence of interactive, user-centric approaches in AI research [14, 15] is in line with the neurosymbolic approach proposed in this paper. In [10] Concept Embedding Models (CEMs) are introduced to try and capture meaningful semantics and to facilitate test-time concept interventions, addressing the accuracy-explainability trade-off. Similarly, [11] demonstrate that GlanceNets can offer interpretable, leak-proof models that maintain task accuracy while improving real-world applicability. [12] leverage language models to guide concept bottlenecks in image classification to enhance interpretability. [13] focus on Label-Free Concept Bottleneck Models, designed for clarity and interpretability in label-scarce scenarios. All these advancements resonate with the objectives of our proposal for alignment using the neurosymbolic cycle. Synergies between these concept-based models and our neurosymbolic approach are expected to open new avenues for AI systems that are not only accurate and efficient but also align with human cognitive processes and ethical standards.

Complementing these developments, [14] introduced the Deep Concept Reasoner (DCR) and ProtoPDebug for concept-level debugging in Part-prototype Networks (ProtoPNets) to enable user-driven model refinement. [15] developed this theme further with few-shot semantic segmentation, based on part-aware prototypes, enhancing prototype representation with both labelled and unlabelled data through interaction. Their user-driven prototype enhancement is in harmony with our method's focus on using logic as a communication bridge between humans and AI. This underscores our shared objective of making AI systems more interpretable and aligned with human understanding, where logic serves not just for internal reasoning but as a medium for effective human-AI collaboration, as also advocated in [16].

Concerning the recent debate around reasoning in deep networks, [17] aims to integrate neural learning with symbolic knowledge representation and logical reasoning for enhanced model interpretability. The debate on the reasoning capabilities of neural networks is much older though; see [18] for an account of the literature on the issue. Although the focus of this paper is on alignment with human values, rather than reasoning per se, our adoption of a logic language lends itself well to the study of reasoning, including comparison of reasoning taking place inside the neural network with reasoning at the symbolic level. Recent findings by [19] pointed to reasoning shortcuts in neurosymbolic models. Reasoning shortcuts occur when models learn concepts that align with logical specifications but diverge from expected human interpretations. Our approach acknowledges and seeks to address this problem by enabling reasoning at the symbolic level to fix reasoning shortcuts at the neural network. Using LTN, logic statements are associated with satisfiability values. Logically-sound reasoning taking place at the symbolic level can therefore be aligned with learned concepts measured against satisfiability values in the network model. This attention to semantic alignment as part of a human-in-the-loop approach is expected to mitigate the risks of reasoning shortcuts by making the model's interpretations and explanations correspond as much as possible to human interpretations, according to the satisfiability metric.

3. Logic as Lingua Franca

To obtain intuitive, human-like explanations, knowledge must be represented at an adequately high level of abstraction, with an ability to drill down into deeper explanations as the need arises, for example as in the case of

a child's sequence of *why* questions. While the predominant, statistics-based explanations and visualizations are effective for debugging, logic-based explanations are precise, require an abstract description of knowledge, and can use such knowledge for reasoning to help elucidate further the explanations [20]. Logic-based explanations can be measured and offer a formal language for evaluation and comparison of results [21]. In [16], symbolic representations are proposed as a means to bridge the gap between human and artificial intelligence. At a minimum, it is argued, symbols should be made available to aid communication, enhance explainability and to explore the importance of advice in human-AI interaction. As well as emphasizing the role of symbols in enabling AI systems to engage meaningfully with humans, in this paper we go one step further and argue for the use of a formal symbolic logic language. In other words, the simple use of symbols in the input and output of a neural network, as in the case of the textual input and output of a large language model, is deemed to be insufficient because it uses an informal and ambiguous natural language.

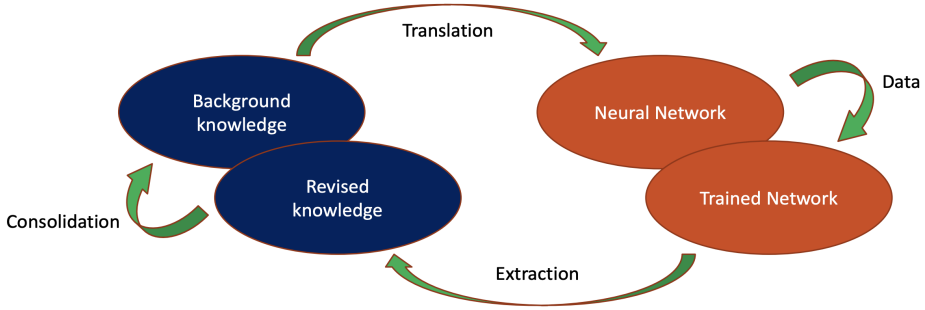

Fig. 1. Illustration of the neurosymbolic cycle [22]: knowledge extraction will be carried out by querying a deep network interactively and aligning continually, thereby repeatedly applying the neurosymbolic cycle until values are deemed to have been aligned. The neurosymbolic cycle enables a human-in-the-loop approach by offering a common ground for communication and system interaction. Symbolic knowledge representation extracted from the learning system at an adequate level of abstraction allow for knowledge consolidation, comparative evaluations and targeted revision of the neural network model.

To illustrate how the idea of querying a neural network may be applied at different levels to achieve alignment, we note that the instantiation of a neural network in LTN occurs once a grounding is assigned to the network. Using first-order logic syntax, LTN provides the conditions for gaining insight into the network and their respective groundings.

Outputs and inner representations of concepts in any neural network are mapped onto a logical predicate to obtain an explanation viz a viz other predicates (other outputs and inner representations connected via logical connectives: negation, conjunction, disjunction, implication and bi-conditional). A user queries the neural network for symbolic knowledge obtaining a direct interpretation of abstract representations and operations on those representations.

Querying with bounded variables enables analysis on how certain features influence outcomes. For instance, one could query whether variables under certain risk conditions share similar outcomes. Specifically, the query ensures that for any individuals or entities x and y from different groups (protected and unprotected) but within the same risk category, their outcomes under function G should be equivalent, reflecting fair treatment across groups. Formally, the query is articulated as: $\forall x, y : \text{Risk}(x) = \text{Risk}(y) \Rightarrow (G(x) \leftrightarrow G(y))$ This query mandates that the model's outcomes, $G(x)$ and $G(y)$, should not only consider the individual's or entity's risk level but also ensure that the outcomes are consistent and non-discriminatory across different demographic or group partitions that share similar risk profiles.

While significant contributions have been made to extracting concepts from inner representations, here we emphasise the importance of understanding the inherent mechanisms at play. A querying method that provides information about conceptual relationships, such as the conjunction of literals in logic, is what is needed to promote alignment of human and AI system. It is through the logical operators that we seek to gain an understanding of the types of relationships that the concepts might form in the neural network. In LTN, first-order logic is used as a rich formalization of the complex communication process. This is not to say that the process of formulating queries in logic is an easy task. We are well aware of the complications that may emerge in a given application, but with a rich formalization

the process can be iterated and evaluated systematically. The use of first-order logic should be accepted as being a relevant approximation of the processes at play.

4. Interactive Alignment and Concept Grounding

Before discussing the practical framework in which the above-mentioned alignment may be carried out, let us clarify the terminology used in the context of neurosymbolic alignment. In this paper, *grounding* refers to the process of mapping abstract symbols and concepts in AI algorithms to real-world data and phenomena. Unlike traditional symbol grounding, which often refers to linguistic symbols, our approach emphasizes grounding in an interactive AI system, where symbols are connected to the AI system's perception and decision-making processes.

In the context of a neurosymbolic AI model, *concepts* refer to high-level, interpretable units of knowledge derived from data. These concepts are represented symbolically and serve as a bridge between the deep learning network and human-understandable reasoning.

Explanations in our model pertain to the methods and processes by which the AI system communicates its decision-making process or reasoning to a human user. Explanations are crucial for measuring whether the AI system's actions are aligned with human values and understanding, including that of non-expert users.

Semantic representation space in our work refers to the multi-dimensional space in which concepts and their relationships are represented. This space is shaped by both the symbolic logic rules and the patterns learned from data by the neural network.

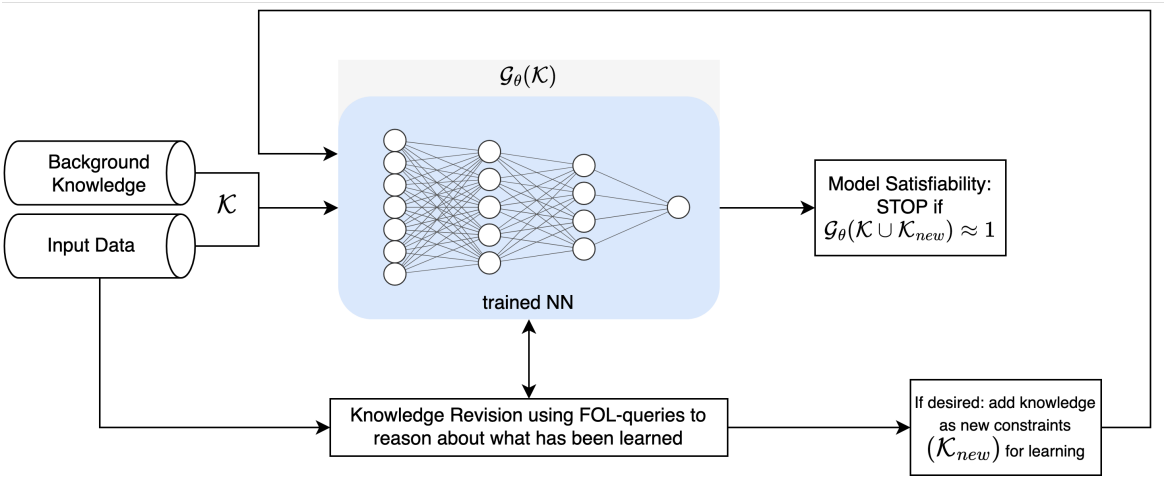

Fig. 2. Illustration of the LTN interactive-learning pipeline: knowledge revision is carried out by querying a deep neural network interactively and learning continually, thus applying the neural-symbolic cycle multiple times. The incorporation of knowledge into the training of the deep network, articulated through first-order logic constraints, has been empirically demonstrated to enhance system fairness [23]. This enhancement is measured in direct comparison with prevailing standard methods resulting from alignment with the decision-making expectations of domain experts.

5. Alignment using the Logic Tensor Network framework

We shall illustrate how a user may query a neural network for symbolic knowledge so that a direct interpretation of abstract representations and their logical operations become available. The goal is to explain and possibly revise the network's decision-making process. It is model-agnostic. Furthermore, to ensure that the model can adapt to

the complexity of tasks in the same way as popularised by current ML, we retain the advantages of gradient-based end-to-end learning. We will need to ensure that the model can be queried with human-interpretable operations at an adequate level of abstraction, so as to guarantee that the common *communication layer* does not create irreconcilable disparities between the symbolic (discrete) and neural (continuous) representations. Logic will provide the formal semantics required for this. Although the use of logic may be seen as a barrier, we argue that it is required to formalise the user interactions. Difficulties that may exist around the use of logic can be ameliorated with the use of *wrappers* to help formulate the logical queries [24].

The building block will be the usual logic operators. The logic operators *connect* the symbol representations in the usual way. Symbols are tangible references that will be used to denote abstract concepts emerging through learning of model-specific, data-driven representations. These abstract concepts will be derived from the trained model, giving rise to explanations that are grounded on the model's representation and logic operations that the ML model has inferred on a given task based on the observed data.

The neurosymbolic framework adopted in this paper is that of Logic Tensor Networks (LTN) as described in [25] and [26]. LTNs function by embedding logic predicates and formulas into a continuous differentiable space, enabling the application of tensor-based learning methods. In simple terms, LTNs allow the AI system to process complex symbolic logic statements using neural network architectures. The primary inputs to an LTN are: (i) symbolic knowledge in the form of logic predicates or formulas; (ii) sub-symbolic data represented by tensors, such as numerical or visual data from various sources, including sensory inputs or databases. Inside an LTN, the symbolic logic statements are translated into a differentiable format that neural networks can process. This is achieved through a grounding process, whereby symbols and logical operators are associated with satisfiability values. The outputs of an LTN are as follows: firstly, it produces predictions or classifications based on the combination of symbolic logic and patterns learned from data; secondly, it provides a logical interpretation for the input data, essentially offering a symbolic explanation or the reasoning behind the AI system's decision-making. This enables the system to learn from data informed by knowledge in a sub-symbolic manner, but hopefully also to reason in a logically-coherent manner, making the decision-making process more interpretable and aligned with human understanding. However, instead of treating the learning of the parameters from data and knowledge as a single process, we emphasise the dynamic and flexible nature of training from data, followed by querying the trained model, followed by consolidating knowledge in the form of constraints for further training, as part of the aforementioned cycle with stopping criteria defined by the user. We make LTN iterative by saving the parameterization learned at each cycle in our implementation of LTN. This does not require the use of any specific, pre-defined neural network architecture.

LTN implements a many-valued first-order logic (FOL) language $\mathcal{L}$ into deep networks. The syntax of LTN is that of FOL, with formulas consisting of predicate symbols, here denoting concepts, and the connectives: negation ($\neg$), conjunction ($\wedge$), disjunction ($\vee$), implication ($\rightarrow$) and bi-conditional ($\leftrightarrow$), as well as universal ($\forall$) and existential ($\exists$) quantification. To emphasize that symbols are interpreted according to their grounding onto real numbers, LTN uses the term *grounding*, denoted by $\mathcal{G}$, in place of logical *interpretation*. Here, we are specifically interested in the grounding of predicates which make up the symbols that refer to the abstract concepts which will form the basis of our model explanation.

In our approach, specific concepts are *probed* and grounded into the LTN framework utilising resources such as the Broden dataset [27], a comprehensive collection that combines a wide array of visual concepts with lexical terms. The list of concepts in this dataset may be incomplete, but it provides rich visual annotations that can aid grounding of abstract concepts into tangible, observable phenomena. By leveraging the Broden dataset, we seek to ensure that the concepts learned by the model are not only semantically meaningful but also visually representative. Alternative approaches have focused on the discovery of concepts, particularly in the case of visual domains. Techniques such as employing CLIP (Contrastive Language–Image Pre-Training) or utilizing pre-established dictionaries, taxonomies or ontologies have been explored recently to try and systematically uncover and categorize concepts [28], thus increasing the applicability of the method proposed here for interactive alignment and interpretability.

Predicates are implemented as mappings onto the interval $[0, 1]$ representing the predicate's degree of truth given the input. In order to allow for gradient-based learning with logical constraints, the logical formulas are made differentiable. This is done by defining the connectives according to fuzzy logic connectives approximated via t-norms, t-conorms and fuzzy implication and negation. These are mathematical operations that satisfy a set of logical properties. In the same manner, quantification such as $\forall$ (for all) and $\exists$ (there exists), and aggregation of

logical formulas into a set (a knowledge-base) are defined using generalised mean. For an extensive explanation of LTN, we refer the reader to [26].

Our adaptation of the LTN framework applies after training. At this point, specific outputs and internal representations of any neural network are mapped onto a predicate P to obtain an explanation for P with respect to other predicates (other outputs and internal representations), defined by the user in the form of a FOL formula. Given the corresponding truth-value of the logic formula establishing a relevant relationship among the predicates, the user may wish to impose additional constraints onto the network through the modification of the knowledge-base, based also on the user's understanding of the relationship between this and other possible formulas.

Assuming that all predicate mappings onto a network are learnable, they all will depend on a set of parameters $\boldsymbol{\theta}$. The initial knowledge-base consists of a (possibly empty) set of logical rules, referred to as ϕ, which entails all predicate mappings. Since the grounding of a rule $\mathcal{G}_\theta(\phi)$ denotes the degree of truth of ϕ, one natural training signal is the degree of truth of all the rules, including mappings in the knowledge-base $\mathcal{K}$. The objective function is built therefore to maximise the satisfiability of all the rules in $\mathcal{K}$, $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \text{Sat}_A(\mathcal{G}_\theta(\mathcal{K}))$, which is subject to an aggregation A of the rules in $\mathcal{K}$.

In our exploration of LTNs for neurosymbolic AI alignment, we recognize that LTNs represent one of several possible methods capable of merging symbolic logic with neural networks. Alternatives, though not the focus of this paper, often employ differential logic-based loss functions, allowing neural networks to adhere to logical constraints while learning from data (e.g. [29, 30]). These methodologies share LTN's goal of blending symbolic reasoning with neural network adaptability, but differ in their specific implementations and optimization. There is now a growing number of approaches in neurosymbolic AI that could be used with our framework, each with its balance of explainability, real-valued logic or probabilistic interpretation, and learning efficacy.

5.1. Illustration of Ethical Alignment using Fairness Constraints

In [23], a method is proposed for acting upon information obtained from XAI approaches to prevent the models from learning unwanted behaviour or biases. The results indicate that integrating the extraction of knowledge from deep networks into the LTN framework and adding tailored fairness constraints for further learning provides a general method for instilling fairness into deep networks. This illustrates how the neurosymbolic method can be used to identify and rectify undesired model behaviour by leveraging XAI methods, in this case SHAP [31]. Moreover, it shows how bias can be identified by utilising the querying mechanism of the proposed framework. The inclusion of symbolic knowledge during the training of deep networks in the form of first-order logic constraints added to the loss function improved quantitative fairness measures while maintaining the performance of the ML system [23] in comparison with existing fairness-specific methods.

One of the key differences between [23] and other XAI-based methods stands out from the fairness experiments. The experiments highlight that simply identifying undesirable model behaviour in certain situations is of limited value. In the fairness experiments, using the LTN framework, once the undesirable model behaviour is identified - in this case, a potential unequal treatment based on gender - model fairness is pursued by adding logic constraints to further training of the model with the objective of minimising bias. To clarify, the fairness constraint in this example addresses gender bias by focusing on different risk classes of two groups: $\mathcal{R}_{M_i}$ (male risk classes) and $\mathcal{R}_{F_i}$ (female risk classes). These indices represent the model's assessment of risk for male and female groups, respectively. They play a key role in ensuring that model learning and decision-making do not favour inadvertently one gender over the other. By incorporating $\mathcal{R}_{M_i}$ and $\mathcal{R}_{F_i}$ in a fairness constraint, the model is actively trained towards a new model that is not only effective but also equitable in its treatment of different gender groups. By creating a continual process, we fully automate the procedure of creating equality groups that are used to instil fairness into the model. For example, we logically constrain equivalence between groups of protected and unprotected classes, whereby for each member (x) of set $\mathcal{R}_{F_i}$ that defaults on credit then a member (y) of set $\mathcal{R}_{M_i}$ should also default and vice-versa. Due to the use of the LTN-based fuzzy logic, satisfiability is determined by aggregation. If aggregation by average is employed (alternative aggregations are possible) then both protected and unprotected groups should default equally on average in the new model.

By conducting experiments across three real-world data sets used to predict income, credit risk and recidivism, it is shown that a neurosymbolic approach can satisfy fairness metrics while maintaining state-of-the-art classification performance [23].

The results are encouraging and indicate that fairness may be achievable in a flexible model-agnostic way. In order for fairness to become a prominent consideration in AI, we must work towards providing practitioners with tools that make it easier to incorporate fairness constraints into existing workflows.

5.2. Illustration of Concept Alignment in Computer Vision

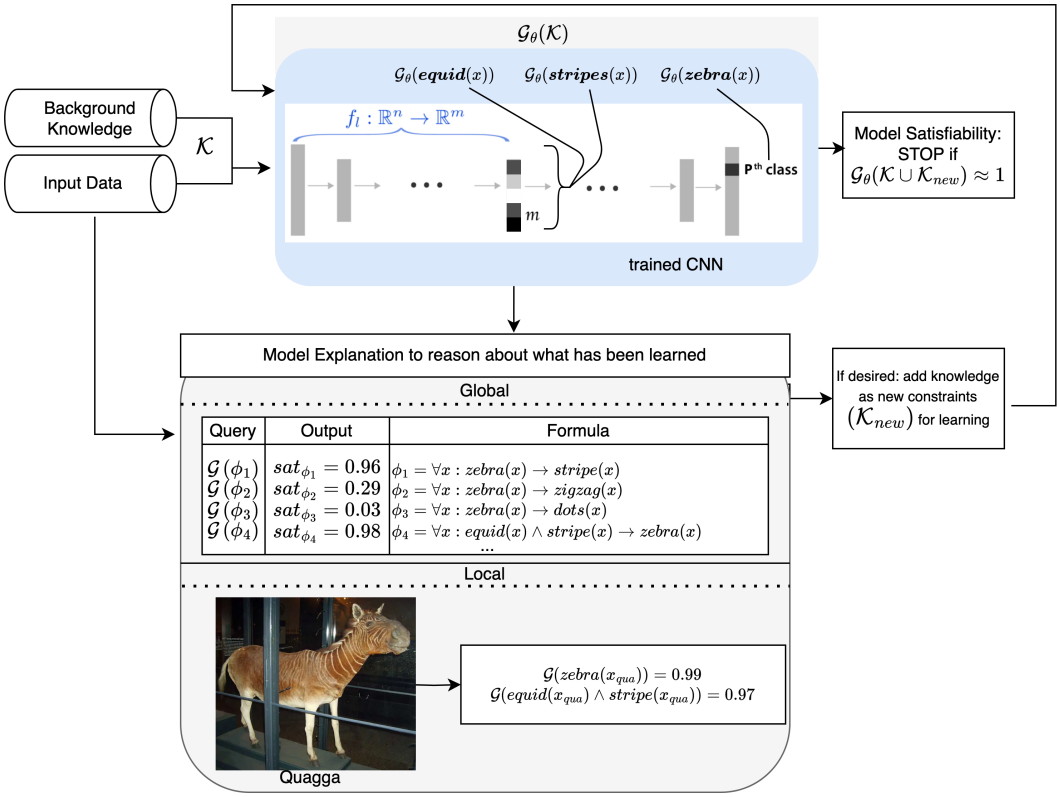

Fig. 3. LTN is queried to produce local explanations (for individual inputs, in this example images) and global explanations (universally-quantified formulas). Reasoning about what has been learned, a user is expected to define the generality of the explanations given the satisfiability of the queries obtained from the trained network. The figure shows some of the queries associated with groundings in the neural network and their satisfiability (sat) levels. Using linear probes to ground the activation patterns of internal representations into the language of LTN, we are able to utilise abstract concepts such as *zebra* and *stripes* as symbols in the logic. After querying, the neural model can be constrained based on a user selection of logical formulas $\mathcal{K}_{new}$ for further training. This iterative process seeks to align the model with user values in the form of symbolic knowledge. In the figure, the Quagga is classified as a zebra initially. The user can change this classification result by adding knowledge into $\mathcal{K}_{new}$ to be satisfied by the new model. Notice that training from data may begin without any knowledge, that is, an empty knowledge-base that can be revised later by querying user-defined concepts and defining constraints, as in the case of the fairness constraints from earlier.

To obtain conceptual explanations that provide comprehensible descriptions of what has been learned, we must ground low-level information into reusable concepts that are present in internal (hidden) representations within the network. Let us illustrate the idea with an example which we have implemented in LTN to explain a Convolutional Neural Network (CNN). We draw inspiration from the TCAV approach [32] but modify it substantially for the implementation in LTN. Consider any neural network that takes as input $\boldsymbol{x} \in \mathbb{R}^n$, which projects onto any layer

l within the network consisting of m neurons, according to a function: $f_l : \mathbb{R}^n \rightarrow \mathbb{R}^m$. In an iterative explanation process, we seek to connect representations inside the network. There is no restriction on which layer l to use, but in a CNN this is generally the layer immediately before the fully-connected layer (i.e. the classifier) [33].

We adapt TCAV [32] to enable users to specify the concepts to be queried at an adequate level of abstraction. It has been shown in [34] in an application to medical imaging that domain-related concepts can be valuable for gaining insight into the decision making process. The approach proposed here extends this idea to allow for complex concept interactions (as defined by the logic) and model retraining using such logical rules as constraints. Using random examples alongside a user-defined set of examples (images) that capture a concept, we form a linear probe at layer l which evaluates the activation values produced by the examples and already known concepts from which further data can be selected for use. In [32], the linear probe at layer l serves as a building block for the Concept Activation Vector used to calculate conceptual sensitivities of inputs and classes. In this paper, we integrate the concept mapping directly into the interactive framework, allowing concepts to be combined into the logic, evaluated using fuzzy logic, and chosen for further alignment of the neural network model. This linear probe works as a classifier for our conceptual grounding, which is then integrated as a logical predicate into the LTN framework. The top part of Figure 3 illustrates the process. At each time that a user wishes to distil a model inference into specific concepts C, they simply select a set P_C of positive examples and a set N of negative examples. The linear probe then serves to distinguish the activation values of the neurons in layer l between $\{f_l(\boldsymbol{x}) : \boldsymbol{x} \in P_C\}$ and $\{f_l(\boldsymbol{y}) : \boldsymbol{y} \in N\}$. This has the advantage of not being bound to pre-existing data or features. Independently of the original task, examples may be collected and any number of user-defined concepts checked (queried) against the network.

For example, we query a GoogLeNet model [35] trained on ImageNet to explain the output class of zebras with respect to user-defined concepts, as illustrated in [32] (bottom part of Figure 3). We extract four different concept descriptions using images from the Broden dataset [27] to dissect the *zebra* classification into the concepts of *stripes*, *dots*, *zigzags* and an abstract representation of the horse-family concept *equidae* (horses, donkeys, zebras and others). We learn the groundings of the activation patterns for the specified concepts from 150 images of each concept and an equal number of negative examples for each class. Subsequently, the truth-value of each query is calculated through fuzzy logic inference using LTN. These queries can be specific to an image (local) or aggregated across the entire set of examples (global). The quantifier $\forall$ is used to aggregate across a set of data points by replacing x with every image available from the dataset thus evaluating the model's behaviour across all available data.[2] The following implication: $\forall x : \mathit{zebra}(x) \rightarrow \mathit{stripe}(x)$, with the symbol $\mathit{stripe}(x)$ being replaced by the corresponding concept grounding in the network, provides an insight into how important the concept $\mathit{stripe}(x)$ is for the CNN's classification output $\mathit{zebra}(x)$ given the set of images x. Furthermore, we can combine several concepts using the logic: $\forall x : \mathit{equid}(x) \wedge \mathit{stripe}(x) \rightarrow \mathit{zebra}(x)$ returns a truth-value of 0.98 across a set of 3000 examples from ImageNet, indicating that the CNN assigns any horse-like object with stripes to the class of zebras. When applying a universal quantifier, the user is able to evaluate the decision making process of the model in general, by examining the concepts on all available data, even if it has not been used for training, thereby producing a global explanation. One example previously unknown to the model is the image of an extinct quagga, an animal characterised by a brown striped coat instead of the black and white pattern of zebras. This example has been selected here to illustrate the potential of local explanations. The model identifies this animal correctly as a member of the equidae family, recognises the stripes on the animal and consequently classifies the image as a zebra, as shown in Figure 3. By utilising the trained linear probes of the activation vectors to ground individual images, we generate local explanations that provide insight into why a particular image might be classified in a certain way by the model. Upon identifying a potential undesired behaviour, for example by querying known exceptions, a user can add new rules into the knowledge-base (by adding logical formulas into $\mathcal{K}_{new}$) for further training of the network. In case the specification of quagga and zebra is to be changed, an alternative inference process can be imposed on the CNN model. The quagga is currently considered by the CNN to be a subspecies of zebra. Assuming that the user decides

[2] This examples illustrates a relevant distinction between the grounding of a statement of the form $\forall x : P(x)$ which applies to all the available examples, and the statement in logic itself, which applies to a possibly infinite set. This distinction which is now made explicit in LTN through the separation of data and knowledge, is at the heart of the definition of the reasoning that is possible and takes place within the network and the reasoning that takes places symbolically. We conjecture that only the reasoning that takes place symbolically in FOL can produce extrapolation beyond what the network can reason about through generalization. A proof of this conjecture should confirm the need for neurosymbolic AI.

to change this, as an example, let us consider introducing concept probes $bw(x)$ for *black and white* objects and $col(x)$ for *colourful* objects, and let us assume that these concepts are to be regarded as mutually exclusive. Adding the following rule to $\mathcal{K}_{new}$ as a new constraint to be satisfied by further learning forces the neural model to classify non-*black and white* objects as non-zebras: $\phi_5 = \forall x : equid(x) \wedge stripe(x) \wedge \neg bw(x) \rightarrow \neg zebra(x)$.[3] Before further training, ϕ_5 exhibits a low *sat*-level of $sat_{\phi_5} = 0.09$ as the model classifies all objects associated with $equid(x)$ and $stripe(x)$ in the *zebra* class. By retraining for only five iterations, the *sat*-level increases to $sat_{\phi_5} = 0.94$. Therefore, the example image of the quagga is no longer inferred to be in the zebra class with $\mathcal{G}(zebra(x_{qua})) = 0.08$, where x_{qua} denotes the image of a quagga. The neural model nevertheless identifies correctly the equidae and stripe concepts in the quagga, with $\mathcal{G}(equid(x_{qua}) \wedge stripe(x_{qua})) = 0.97$. It should be noted that the explanation itself does not affect the performance of the model. Thus, prior to revising $\mathcal{K}$, the behaviour of the model remains unchanged due to the use of linear probes that solely interpret activation patterns.

6. Conclusion and Future Work

This paper aimed to capture and propel the momentum of the field of AI towards more ethical, transparent and human-centric AI. It bridges the gap between various approaches, synthesizing the strengths of incremental interactive learning, concept-based reasoning, and neural-symbolic integration. This synthesis is crucial for the development of AI systems that are not only technically advanced but also ethically sound, socially responsible, and aligned with human values. The field is at a pivotal juncture where such integrative approaches can lead to significant advancements in AI that are beneficial to society.

As Machine Learning methods are becoming more widely used, it will be essential to provide explanations at various levels of abstraction and complexity. As part of this, it will be important to provide flexible explanations enabled by the central idea of querying a neural network symbolically. Furthermore, granting users the ability to influence the decision-making process will help drive wider adoption across a range of relevant applications.

In this paper, we have presented recent efforts at steering AI systems towards alignment with a specific focus on incorporating fairness and providing explanations as part of a neurosymbolic approach. The connection between fairness and explanation in our framework is fundamental to achieving value alignment in AI systems. Fairness ensures that the model's decisions do not unjustly favour or disadvantage any group or individual, adhering to ethical standards and societal values. Explanation, on the other hand, provides transparency and understanding of the model's decision-making process, allowing users to trust and effectively interact with the AI system. However, these elements alone are not sufficient to guarantee value alignment. True value alignment in AI requires a holistic approach that encompasses fairness and explainability together as part of the neurosymbolic cycle, but also other crucial aspects such as respect for privacy, non-maleficence, and autonomy, which is left as future work. Our framework integrates fairness and explainability by continuously adapting and revising its decisions in light of new data and user feedback, hopefully aligning the AI's actions with human values dynamically over time.

In the proposed framework, neurosymbolic integration acts as a common layer of communication in which different levels of abstraction can be utilised to exchange information from the model to its human counterpart and vice versa. The alignment of model and human reasoning with the objective of facilitating better understanding of AI systems should enable us to increase the trust we have in AI models. To achieve this, it is necessary to address both the grounding and the relational components of symbolic systems. An integration of knowledge and data into a neural network, with the ability to extract post-hoc information and perform continual enhancement based on knowledge revision, has been proposed as a bottom-up neurosymbolic approach to human-AI alignment. Our approach contrasts with the current trend of training very large models by gobbling up all of the data in the internet, the so-called *scale is all you need* approach. In addition to the concerns around energy efficiency, this approach seems to create the difficult problem of post-hoc alignment, along with the associated ethical problems already mentioned. Hence, the question of how to fix large language models (LLM) using a neural-symbolic approach or develop more modular models with the help of data and knowledge is a complex one. Further research will be needed to explore

[3]Notice that the satisfiability of this rule should be the same as that of $\forall x : equid(x) \wedge stripe(x) \wedge col(x) \rightarrow \neg zebra(x)$, as we apply a $softmax$ function to mutually exclusive concept probes; in this case $\forall x : col(x) \leftrightarrow \neg bw(x)$.

and evaluate the adaptation of the neural-symbolic cycle to LLMs to enhance stability and continuous learning. Alternatively, the development of new models that incorporate modular design and knowledge may offer a pathway to overcome the limitations of current LLMs. The observed inconsistencies in GPT-4's performance and the challenges of catastrophic forgetting present opportunities for innovative approaches. Further research and experimentation are needed to determine the most effective strategies for enhancing the stability, explainability, and continuous learning capabilities of LLMs.

The neurosymbolic approach shows promise for developing AI systems that are aligned with human goals, values, and expectations. By combining neural and symbolic techniques, neurosymbolic agents could be more transparent, consistently interpretable, and amenable to human oversight. Further research and practical experimentation at larger scale is needed to fully realize the potential of neurosymbolic AI for safe and beneficial AI systems. The pursuit of developing models capable of yielding generalizable concepts constitutes an ongoing research endeavor. As the complexity and size of models escalate, there is emerging evidence that these models are proficient in constructing increasingly intricate and transferable representations. Within this context, the consideration of data multimodality is perceived as a pivotal advancement. The method delineated herein is posited to be particularly applicable to scenarios involving multimodality, such as the combination of images and text or other sensory input. Training joint-embeddings on images and text, for example, is expected to help bridge the gap between discrete and continuous representations and underscore the relevance and contribution of neurosymbolic systems to AI.

References

[1] D.M. Ziegler, N. Stiennon, J. Wu, T.B. Brown, A. Radford, D. Amodei, P. Christiano and G. Irving, Fine-Tuning Language Models from Human Preferences., *CoRR* **abs/1909.08593** (2019). https://arxiv.org/pdf/1909.08593.

[2] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L.E. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike and R.J. Lowe, Training language models to follow instructions with human feedback, *ArXiv* (2022). https://www.semanticscholar.org/paper/d766bffc357127e0dc86dd69561d5aeb520d6f4c.

[3] S. Casper, X. Davies, C. Shi, T.K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, T. Wang, S. Marks, C.-R. Segerie, M. Carroll, A. Peng, P. Christoffersen, M. Damani, S. Slocum, U. Anwar, A. Siththaranjan, M. Nadeau, E.J. Michaud, J. Pfau, D. Krasheninnikov, X. Chen, L. Langosco, P. Hase, E. Bıyık, A. Dragan, D. Krueger, D. Sadigh and D. Hadfield-Menell, Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, 2023.

[4] D.L. Silver and T.M. Mitchell, The Roles of Symbols in Neural-based AI: They are Not What You Think!, in: *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning, La Certosa di Pontignano, Siena, Italy, July 3-5, 2023*, A.S. d'Avila Garcez, T.R. Besold, M. Gori and E. Jiménez-Ruiz, eds, CEUR Workshop Proceedings, Vol. 3432, CEUR-WS.org, 2023, pp. 420–421. https://ceur-ws.org/Vol-3432/paper40.pdf.

[5] G. Futia and A. Vetrò, On the integration of knowledge graphs into deep learning models for a more comprehensible AI-Three challenges for future research, 2020. ISSN 20782489. doi:10.3390/info11020122.

[6] K.H. Ngan, J. Phelan, E. Mansouri-Benssassi, J. Townsend and A.S. d'Avila Garcez, Closing the Neural-Symbolic Cycle: Knowledge Extraction, User Intervention and Distillation from Convolutional Neural Networks, in: *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning, La Certosa di Pontignano, Siena, Italy, July 3-5, 2023*, A.S. d'Avila Garcez, T.R. Besold, M. Gori and E. Jiménez-Ruiz, eds, CEUR Workshop Proceedings, Vol. 3432, CEUR-WS.org, 2023, pp. 19–43. https://ceur-ws.org/Vol-3432/paper3.pdf.

[7] P.W. Koh, T. Nguyen, Y.S. Tang, S. Mussmann, E. Pierson, B. Kim and P. Liang, Concept bottleneck models, in: *International conference on machine learning*, PMLR, 2020, pp. 5338–5348.

[8] S. Teso, K. Kersting et al., Explanatory interactive machine learning, in: *AIES 2019-Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, Association for Computing Machinery, Inc, 2019, pp. 239–245.

[9] G. Schwalbe, Concept Embedding Analysis: A Review, 2022.

[10] M.E. Zarlenga, P. Barbiero, G. Ciravegna, G. Marra, F. Giannini, M. Diligenti, Z. Shams, F. Precioso, S. Melacci, A. Weller, P. Lio and M. Jamnik, Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off, 2022.

[11] E. Marconato, A. Passerini and S. Teso, Glancenets: Interpretable, leak-proof concept-based models, *Advances in Neural Information Processing Systems* **35** (2022), 21212–21227.

[12] C. Yang, A. Rangarajan and S. Ranka, Global Model Interpretation via Recursive Partitioning (2018).

[13] T. Oikarinen, S. Das, L.M. Nguyen and T.-W. Weng, Label-Free Concept Bottleneck Models, 2023.

[14] W. Stammer, P. Schramowski and K. Kersting, Right for the Right Concept: Revising Neuro-Symbolic Concepts by Interacting with their Explanations, 2021, pp. 3618–3628. doi:10.1109/CVPR46437.2021.00362.

[15] A. Bontempelli, S. Teso, F. Giunchiglia and A. Passerini, Concept-level Debugging of Part-Prototype Networks (2022). doi:10.48550/arXiv.2205.15769.

[16] 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, *Proceedings of the AAAI Conference on Artificial Intelligence* **36**(11) (2022), 12262–12267. doi:10.1609/aaai.v36i11.21488. https://ojs.aaai.org/index.php/AAAI/article/view/21488.

[17] P. Barbiero, G. Ciravegna, F. Giannini, M.E. Zarlenga, L.C. Magister, A. Tonda, P. Lio', F. Precioso, M. Jamnik and G. Marra, Interpretable Neural-Symbolic Concept Reasoning, 2023.

[18] A.S. d'Avila Garcez and L. Lamb, Neurosymbolic AI: the 3rd wave, *Artificial Intelligence Review* (2020), 1–20. https://api.semanticscholar.org/CorpusID:228083996.

[19] E. Marconato, S. Teso, A. Vergari and A. Passerini, Not All Neuro-Symbolic Concepts Are Created Equal: Analysis and Mitigation of Reasoning Shortcuts, *arXiv preprint arXiv:2305.19951* (2023).

[20] R. Confalonieri, L. Coba, B. Wagner and T.R. Besold, A historical perspective of explainable Artificial Intelligence, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2021). doi:10.1002/widm.1391.

[21] J. Townsend, M.-B. Esma, H.N. Kwun and A. d'Avila Garcez, Chapter 16. Discovering Visual Concepts and Rules in Convolutional Neural Networks, in: *Frontiers in Artificial Intelligence and Applications*, Volume 369 edn, 2023, pp. 337–372. doi:10.3233/FAIA230148.

[22] A.d. Garcez, K.B. Broda and D.M. Gabbay, *Neural-Symbolic Learning Systems*, Perspectives in Neural Computing, Springer London, London, 2002, p. 275. ISBN 978-1-85233-512-0. doi:10.1007/978-1-4471-0211-3.

[23] B. Wagner and A.d. Garcez, Neural-Symbolic Integration for Fairness in AI, in: *AAAI Spring Symposium AAAI-MAKE*, 2021. http://ceur-ws.org/Vol-2846/paper5.pdf.

[24] H. Singh, M. Aggarwal and B. Krishnamurthy, Exploring Neural Models for Parsing Natural Language into First-Order Logic, *CoRR* (2020). https://arxiv.org/abs/2002.06544.

[25] L. Serafini and A.d. Garcez, Logic tensor networks: Deep learning and logical reasoning from data and knowledge, *arXiv preprint arXiv:1606.04422* (2016).

[26] S. Badreddine, A.d. Garcez, L. Serafini and M. Spranger, Logic Tensor Networks (2020). http://arxiv.org/abs/2012.13635.

[27] D. Bau, B. Zhou, A. Khosla, A. Oliva and A. Torralba, Network dissection: Quantifying interpretability of deep visual representations, in: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. ISBN 9781538604571. doi:10.1109/CVPR.2017.354.

[28] A. Radford, J. Wook, K. Chris, H. Aditya, R. Gabriel, G. Sandhini, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, arXiv:2103.00020, 2021.

[29] F. Giannini, M. Diligenti, M. Maggini, M. Gori and G. Marra, T-norms driven loss functions for machine learning, *Applied Intelligence* (2023), 1–15.

[30] K. Ahmed, K.-W. Chang and G.V. den Broeck, A Pseudo-Semantic Loss for Deep Generative Models with Logical Constraints, in: *NeurIPS*, 2023.

[31] S.M. Lundberg and S.I. Lee, A unified approach to interpreting model predictions, in: *Advances in Neural Information Processing Systems*, 2017. ISSN 10495258.

[32] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas and R. Sayres, Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV), in: *35th International Conference on Machine Learning, ICML 2018*, 2018. ISBN 9781510867963.

[33] S. Odense and A. d'Avila Garcez, Layerwise Knowledge Extraction from Deep Convolutional Networks, *CoRR* **abs/2003.0** (2020). https://arxiv.org/abs/2003.09000.

[34] M. Graziani, V. Andrearczyk and H. Müller, Regression Concept Vectors for Bidirectional Explanations in Histopathology, in: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, pp. 124–132. ISSN 16113349. ISBN 9783030026271.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, Going deeper with convolutions, in: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015. ISSN 10636919. ISBN 9781467369640. doi:10.1109/CVPR.2015.7298594.