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 explanations. 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 a Convolutional Neural Network. 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 fairness constraint.

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

1. Introduction

As AI systems become increasingly popular and influential in various domains, there is a growing concern regarding their potential misalignment with human values and goals. To address these concerns, researchers have advocated the development of more transparent and interpretable AI models, capable of providing meaningful insights to the users about the model's internal processes. Users, in particular system developers, may also wish to have the ability to influence or control the system's decision-making processes. In response to this, in this paper we propose a neurosymbolic integration approach that leverages symbolic reasoning and deep learning to achieve both explainability and better alignment. Specifically, we focus on a hybrid system where users and domain experts can interact with the system during its training phase, allowing users to make sense of how the system operates and to modify its behaviour as needed. We argue that the neurosymbolic approach offers a pathway to the development of AI systems that are more transparent and aligned with human goals. We will illustrate how symbolic logic can be used to query a deep network to close the neurosymbolic cycle (Figure 1) using a practical application in computer vision.

In order to facilitate interaction between AI systems and humans, we draw inspiration from previous research on the use of language in the communication between cognitive agents [1]. Language in this context refers to a means of representing and communicating about the world. In [2], the author draws attention to the problem known as the symbol grounding problem, in which natural and artificial cognitive agents must develop an intrinsic connection between their symbolic representations and some referents in the external world. In this process, individuals may represent external referents in a conceptual manner and use them as grounding for symbolic representations. The development of these internal representations occurs as a result of interactions with entities in the external world.

*Corresponding author. E-mail: Benedikt.Wagner@city.ac.uk.

Learning how to categorise allows us to form discrete and useful concepts of our environment. Based on the following chain of representations and entities, the cognitive mechanism underlying the grounding of physical symbols is formed [1]:

$$\text{external entities} \Leftrightarrow \text{internal representations} \Leftrightarrow \text{symbols}.$$

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 represent and act in the world [3].

During learning, the most distinguishing characteristic of this type of perception is the requirement to warp the perceived into a similarity space of internal categorisations [1]. For the purposes of this process, we believe that a neural network is the most suitable method of learning transformations that may be capable of producing semantic representation spaces. Thus, we are interested in focusing on the bidirectional translation of neural networks into concepts. In what follows, we illustrate how we approach this problem using the proposed framework.

In the area of knowledge extraction and model explanation, there exists a tension between the desire for logical precision and the necessity of practicality. While precision may be desirable, it is often impractical, particularly in the context of symbol grounding. Within the domain of computer vision, this challenge becomes especially salient. Explanations rooted in logical relations of pixel values may offer precision, but they also risk succumbing to excessive complexity, rendering them unusable in practice. Instead, 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 operate at this level of abstraction, yielding a more intuitive experience for users.

Futia et. al [4] 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 applications 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* [4]. 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 setting neurosymbolic integration should allow us to overcome the static nature of the current ML paradigm. Explainability methods of today do not provide an ability to act on extracted information. Upon finding an undesired property of the system, the only way to influence the model is to retrain it until a good model is found. However, retraining as a process may be unguided and can only be influenced indirectly by the explanation, normally through the collection of additional data. The result is that many XAI methods become limited in their usefulness, with retraining commonly resulting in catastrophic forgetting of previously acquired information. Moreover, [5] provided an empirical evaluation of the performance of large language model (LLM) GPT-4 over time, uncovering significant variations. For instance, GPT-4's accuracy in identifying prime numbers exhibited a marked decline from 97.6% to 2.4% between March and June 2023. This temporal inconsistency raises critical questions about the stability, robustness, and direction of improvement of LLMs. Connecting these insights, the observed decline in GPT-4's performance might be indicative of the broader challenges of retraining and catastrophic forgetting. A neural-symbolic approach is proposed here as an alternative to addressing the inconsistencies observed by [5], one that is based instead on the use of a formal logical language.

If one wishes to obtain intuitive, human-like explanations then the alignment must take place at a higher 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 low-level and statistical explanations are effective for debugging, logic-based explanations are precise, require a description at a more abstract level of representation, and can use such knowledge representation for reasoning which can elucidate further the explanation using higher-level concepts [6]. Logic-based explanations can be measured and offer a common language for evaluation and comparison of results [7].

One of the core premises of the theory of communication is that communication is primarily symbolic [9]. According to symbolic interactionism, humans interpret and assign meaning to events through a set of symbols that are interconnected. Over the years, much research has been conducted on communication based on these principles,

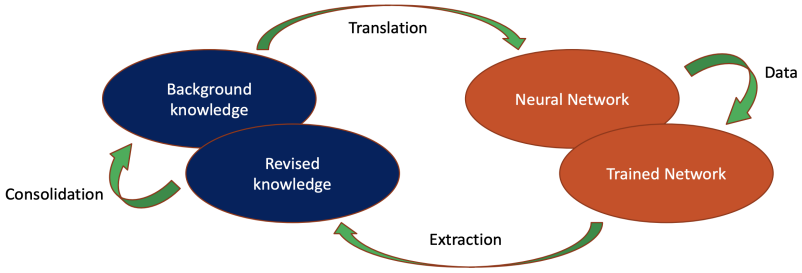

Fig. 1. Illustration of the neurosymbolic cycle [8]: knowledge extraction will be carried out by querying a deep network interactively and aligning continually, thereby repeatedly applying the neurosymbolic cycle until values are deemed to have been aligned. The neurosymbolic cycle enables a human-in-the-loop approach by offering a common ground for communication and system interaction. Symbolic knowledge representations extracted from the learning system at an adequate level of abstraction allow for knowledge consolidation, comparative evaluations and targeted revision of the neural network models.

which can provide a broad framework for all forms of communication, including non-verbal communication. Furthermore, processes that provide information about concept interaction and relationships, such as the conjunction of literals in logic, need to be captured by the framework. It is through the logical operators that we seek to gain an understanding of the types of relations and abstractions that the concepts might form. This should be accepted as being an approximation, albeit a computationally relevant one with the use of first-order logic (FOL), as it may not be possible to capture the complexities of all interactions in their entirety using FOL.

Let us illustrate the value of the combination of neural networks and FOL. A neural network's distributed representation is by definition grounded on its set of examples. It is reasonable to assume that reasoning capabilities with higher-order concepts and knowledge representation may emerge from training such deep neural networks. In a trained network, however, to show that a certain property holds true for all (possibly infinite) inputs requires constructing (and listing) all cases; the neural network is an efficient, grounded computational constructive model. When a description of the form $\forall x P(x)$ is acquired, though, denoting all the inputs with property P, an extrapolation to infinite domains takes place, one that is not grounded and does not require listing all cases. It is this representational plurality to and from an efficient computational model and its rich descriptions and references that we seek to explore. On this note, there is nothing stopping us from representing $\forall x P(x)$ in the network itself but in a different, rather more localist representation. In this sense, a symbolic representation emerges to enable communication and itself influences representation, allowing for introspection, extrapolation and reasoning by analogy, all key elements of general AI systems.

After starting this paper with an introduction to the imperative for transparent alignment within AI models and an exposition of the neurosymbolic integration approach. Subsequently, we investigate the intricate relationships between entities, representations, and symbols. We then delineate a neurosymbolic methodology designed to perpetually align a neural network with human expectations, utilizing symbols and logic as a lingua franca. In the next section, we provide a succinct example within the domain of computer vision to elucidate the practical implementation of this approach.

2. Interactive Alignment and Concept Grounding

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 approach seeks to explain and possibly revise a decision-making process post-hoc. 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 semantic precision required. Although the use of logic may be seen as a barrier initially, we argue that it is required to

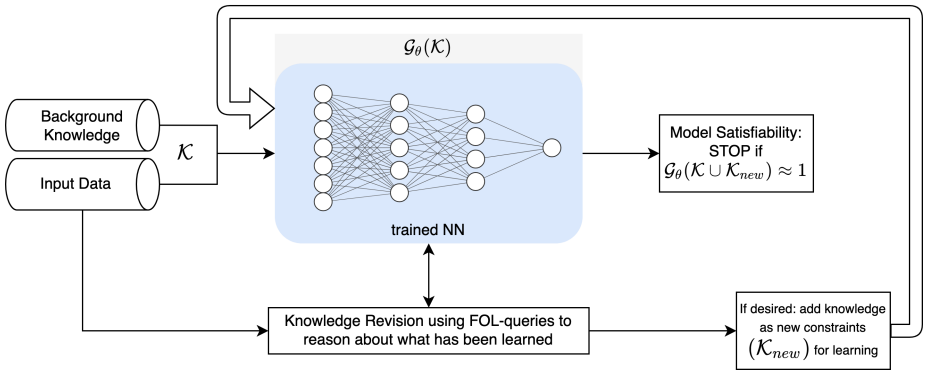

Fig. 2. Illustration of the LTN interactive-learning pipeline: knowledge revision will be carried out by querying a deep network interactively and learning continually, thus applying the neural-symbolic cycle multiple times. The incorporation of knowledge into the training of the deep network, articulated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

formalise the user interactions. Difficulties that may exist around the use of logic can be ameliorated with the use of *wrappers* to formulate logical queries in natural language [10]. The core building block will be the usual logical operators. The goal is to provide an intuition into the operations that the ML model has inferred on a given task based on the observed data. The logical operators *connect* the symbol representations in the usual way. Symbols are tangible references which will be used to denote abstract concepts that emerge 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 inherent operations.

The neurosymbolic framework adopted in this paper is that of Logic Tensor Networks (LTN) as described in [11] and [12]. 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 parametrization learned at each cycle in our implementation of LTN, while not requiring the use of any specific neural network model or 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 ($\rightarrow$), 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 methodology, specific concepts are predefined, utilising resources such as the Broden dataset. This contrasts with alternative approaches in the field that focus on the discovery of concepts, particularly within visual domains. Techniques such as employing CLIP (Contrastive Language–Image Pre-Training) or utilizing pre-established dictionaries, taxonomies, or ontologies have been explored to systematically uncover and categorize these concepts, contributing to the broader applicability of this method 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 [12].
Our adaptation of the LTN framework can be deployed after training. Here, specific outputs and internal representations of any neural network are mapped onto a predicate P to obtain an explanation for P w.r.t. other predicates (other outputs and internal representations) with the use of the logical connectives. Once a relation among the predicates has been established as a logical rule, one can impose additional constraints onto the network with this rule to seek to understand the relationship between this and other possible rules. If we consider all groundings of mappings of a network to be learnable, they will all 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 $\mathcal{K}$, $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ \mathrm{Sat}_A(\mathcal{G}_\theta(\mathcal{K}))$, which is subject to an aggregation A of the rules in $\mathcal{K}$.

3. Example of Alignment with Fairness Constraints using LTN

In [13], 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 [14]. 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 improves quantitative fairness measures while maintaining the performance of the ML system [13] in comparison with existing fairness-specific methods. One of the key differences between this approach and other XAI-based methods stands out from the experiments. The fairness experiment highlights that simply identifying undesirable model behaviour in certain situations is of limited value. Using the LTN framework, we discuss how to identify potential unequal treatment and illustrate how fairness may be achieved by using constraints in order to effectively minimise biases. By creating a continual process, we fully automate the procedure of creating equality groups that are used to instill fairness into the model. For example, we logically constrain equivalence between groups of protected and unprotected classes, where each member (x) of set $\mathcal{R}_{Fi}$ defaults on credit, i.e. $h(x) = 1$, then a member (y) of set $\mathcal{R}_{Mi}$ should also default, $h(y) = 1$, and vice-versa. Due to our use of the LTN-based fuzzy logic, each 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. By conducting experiments across three real-world data sets used to predict income, credit risk and recidivism, we show that a neurosymbolic approach can satisfy fairness metrics while maintaining state-of-the-art classification performance [13].

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

4. Example of Concept Alignment in Computer Vision

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 [15] 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: $f_l : \mathbb{R}^n \to \mathbb{R}^m$. 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) [16]. We adapt TCAV [15] to enable users to specify the concepts to be queried at an adequate level of abstraction. It has been shown

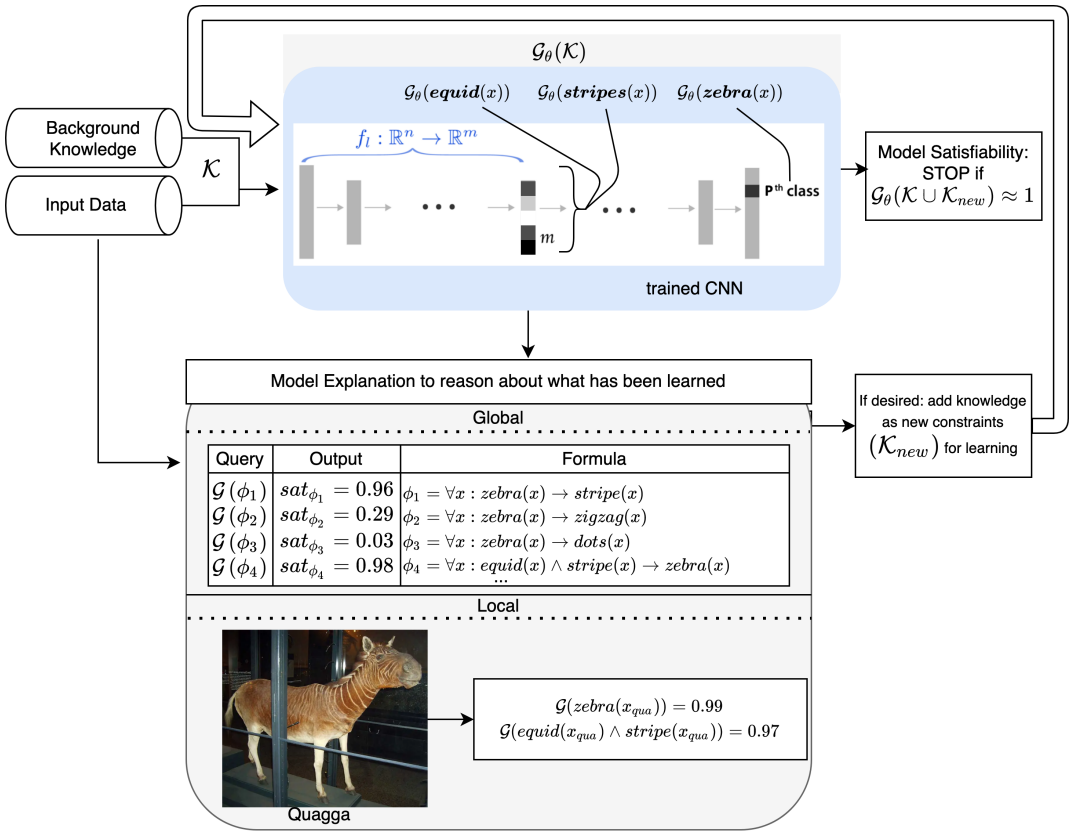

Fig. 3. We produce local explanations (for individual inputs/images) and global explanations (universally-quantified formulas) for the deep learning model by querying it. We then reason about the generality of the explanations given the satisfiability of the queries obtained from the trained network. 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 as symbols in the logic. Following 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 $\mathcal{K}$. In the figure, the Quagga is classified as a zebra. A user's desire to change such classification should trigger the addition of knowledge into $\mathcal{K}_{new}$ informed by the queries to be satisfied by the final trained model. Notice that training from data may begin without any knowledge (an empty knowledge-base) which can be revised later by querying user-defined concepts and constraints, such as the fairness constraints from earlier, deemed as necessary for the network to learn.

in [17] in an application to medical imaging that domain-related concepts can be valuable for gaining insight into the decision making. 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 [15], 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 by pre-existing data or features. Independent of the original task, examples (images) may be collected and

any number of user-defined concepts checked (queried) against the network.
As an example, we query a GoogLeNet model [18] trained on ImageNet to explain the output class of zebras with respect to user-defined concepts, as illustrated in [15] and the bottom part of Figure 3. We extract four different concept descriptions using images from the Broden dataset [19] 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. The following implication: $\forall x : zebra(x) \rightarrow stripe(x)$, with the symbol $stripe(x)$ being replaced by the corresponding concept grounding in the network, provides an insight onto how important the concept $stripe(x)$ is for the CNN's classification output $zebra(x)$ given the set of images x.
Furthermore, we can combine several concepts using the logic: $\forall x : equid(x) \wedge stripe(x) \rightarrow zebra(x)$ 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 extinct quagga, an animal characterised by a brown striped coat instead of the black and white pattern of zebras which has been selected 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 that of 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 according to the model.

Upon identifying 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. Recall that quagga are currently considered by the CNN to be a subspecies of zebra. Assuming that the user decides 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 learning should force the neural model to only classify black and white objects as zebras: $\phi_5 = \forall x : equid(x) \wedge stripe(x) \wedge \neg bw(x) \rightarrow \neg zebra(x)$.[1]
Before further training, ϕ_5 exhibits a low *sat*-level of $sat_{\phi_5} = 0.09$, as the model classifies all objects associated with the $equid(x)$ and the $stripe(x)$ concepts to the *zebra* class regardless of their colour. By retraining for only five iterations, the *sat*-level increases to $sat_{\phi_5} = 0.94$, which indicates that only black and white objects (in conjunction with the *stripe* and *equidae* concepts) are now considered to be zebras. 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.

5. Conclusion and Future Work

As Machine Learning methods are becoming more widely used, it will be essential to provide explanations with varying degrees of abstractness and complexity, as outlined in [4]. As part of this, it may become desireable to

[1] 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 $\forall x : col(x) \leftrightarrow \neg bw(x)$.

present explanations that would enable laymen to to understand the underlying model representations. Moreover, granting users the ability to influence their decision-making process may help drive wider adoption across a range of safety-critical applications.

The integration can be viewed as a paradigm shift allowing for active and interactive engagement between system and user by enabling communication. Therefore, the neurosymbolic integration serves 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 will enable us to increase the trust we have in 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, is proposed as a bottom-up neurosymbolic approach to Human-AI alignment. The question of whether to fix issues around large language models using a neural-symbolic approach or to develop a more modular model with prior linguistic knowledge is complex. Future research may explore the adaptation of the neural-symbolic cycle to LLMs like GPT-4 to enhance stability and continuous learning. Alternatively, the development of new models that incorporate modular design and prior knowledge about language 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, consistent interpretable, and amenable to human oversight. Further research 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's multi-modality is perceived as a pivotal advancement. The method delineated herein is posited to be particularly applicable to scenarios involving multi-modality, underscoring its relevance and potential contribution to the field.

References

[1] A. Cangelosi, The grounding and sharing of symbols, *Pragmatics & Cognition* (2006). doi:10.1075/pc.14.2.08can.

[2] S. Harnad, The symbol grounding problem, *Physica D: Nonlinear Phenomena* (1990). doi:10.1016/0167-2789(90)90087-6.

[3] 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.

[4] 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.

[5] L. Chen, M. Zaharia and J. Zou, How is ChatGPT's behavior changing over time?, 2023.

[6] 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.

[7] 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.

[8] 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.

[9] S.W. Littlejohn, Symbolic interactionism as an approach to the study of human communication, *Quarterly Journal of Speech* (1977). doi:10.1080/00335637709383369.

[10] 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.

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

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

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.