Neurosymbolic Artificial Intelligence -1 (2024) 1–13
DOI 10.3233/NAI-240740
IOS Press
CORRECTED PROOF

Towards a neuro-symbolic cycle for human-centered explainability

Alessandra Mileo
Insight Centre for Data Analytics, Dublin City University, Dublin, Ireland
E-mail: alessandra.mileo@insight-centre.org

Editor: Benedikt Wagner, City, University of London, United Kingdom
Solicited reviews: Qiqi Su, City, University of London, United Kingdom; Two anonymous reviewers

Received 15 August 2023
Revised 17 April 2024
Accepted 2 May 2024

Abstract. Deep learning is being very successful in supporting humans in the interpretation of complex data (such as images and text) for critical decision tasks. However, it still remains difficult for human experts to understand how such results are achieved, due to the "black box" nature of the deep models used. In high-stake decision making scenarios such as the interpretation of medical imaging for diagnostics, such a lack of transparency still hinders the adoption of these techniques in practice. In this position paper we present a conceptual methodology for the design of a neuro-symbolic cycle to address the need for explainability and confidence (including trust) of deep learning models when used to support human experts in high-stake decision making, and we discuss challenges and opportunities in the implementation of such cycle as well as its adoption in real world scenarios. We elaborate on the need to leverage the potential of hybrid artificial intelligence combining neural learning and symbolic reasoning in a human-centered approach to explainability. We advocate that the phases of such a cycle should include i) the extraction of knowledge from a trained network to represent and encode its behaviour, ii) the validation of the extracted knowledge through commonsense and domain knowledge, iii) the generation of explanations for human experts, iv) the ability to map human feedback into the validated representation from i), and v) the injection of some of this knowledge in a non-trained network to enable knowledge-informed representation learning. The holistic combination of causality, expressive logical inference, and representation learning, would result in a seamless integration of (neural) learning and (cognitive) reasoning that makes it possible to retain access to the inherently explainable symbolic representation without losing the power of the deep representation. The involvement of human experts in the design, validation and knowledge injection process is crucial, as the conceptual approach paves the way for a new human–ai paradigm where the human role goes beyond that of labeling data, towards the validation of neural-cognitive knowledge and processes.

Keywords: Explainable AI, neuro-symbolic cycle, graph analysis, rule extraction, human-centric AI

1. Introduction

Thanks to the availability of huge data and computational resources, within the last 10 years Deep Learning has gained popularity and success [31], eliminating the need for complex features' engineering by automatically learning complex data representations of millions of features directly from (millions of) data samples [41]. The ability to automatically classify medical images to support clinicians in early diagnosis is a key application of Deep Learning

in computer vision, given its potential in reducing reporting delays, mitigating human errors and highlighting critical or urgent cases [26]. The issues of explainability and transparency for these models, however, are not systematically addressed, creating a gap between advances in research and impact in clinical practice, and hindering wider adoption [39]. This is reflected not only on diagnostic imaging and health in general, but also more broadly on applications involving environmental issues, societal well-being and fundamental human rights. Such application areas are of crucial importance in the EU proposal for regulation on Artificial Intelligence (COM/2021/206), since an error in the outcome of the model leading to a wrong decision can carry a high cost (high-stake decision making). In order to address the issues of explainability, trust and fairness for the application of Deep Learning in high-stake decision making such as diagnostic imaging, different methods have been proposed to interpret the inner workings of deep learning architectures [50]. Despite advances in this area, the key assumption advocating for Deep Learning versus inherently interpretable models is that there is a trade-off between the accuracy of the model and its interpretability, which is not necessarily the case [38]. A prediction with high accuracy, in fact, is not necessarily trustworthy [52]: when based only on the learning process as it happens in current interpretation methods, model-based interpretability carries the risk of producing incomplete or incorrect explanations [38]. If you consider attribution maps, for example, the portion of an image that is highlighted as responsible for a given classification outcome (e.g. the area around the paw that makes the algorithm classify the dog as a transversal flute or as a husky), does not say anything about why that outcome was produced (is it because of the shape of the paw? The colour? The area around it?), making the explanation incomplete and in some ways misleading.

The current inability of deep representations to include information about cause-effect, compositionality and context is the main gap that needs to be address to enhance understanding and therefore trust in neural models. To this aim, the neuro-symbolic cycle proposed in this position paper will pave the way to close this gap. The first step of the cycle is the ability to map deep representations into a symbolic space via knowledge extraction, so that what the network has learned can be understood and accessed in human terms. From there, two complementary aspects of the proposed neuro-symbolic cycle will be discussed. The first aspect focuses on the ability to revise and correct such knowledge, and inject it back into the learning process (we will call this the neuro-symbolic extraction-injection cycle). The second aspect focuses on the ability to generate explanations for domain experts and use human feedback to enhance and correct cognitive reasoning processes (we will call this the explanation-feedback-control cycle).

The concepts and ideas presented in this paper have general applicability to computer vision tasks, and are not limited to CNN. However, in order to appreciate the impact and significance of the ideas proposed, we will often refer to high-stake decision making tasks such as image classification applied to Medical Image Analysis as a usecase scenario. This is particularly interesting when considering trustworthiness of AI as a decision support tool for experts. Clinical experts, in fact, believe the use of deep learning can speed up the processing and interpretation of radiology data by 20%, reducing errors in diagnosis by approximately 10%.[1] Yet there is still a lack of clinical adoption due to the fact that it remains difficult for humans to understand how such results are achieved due to the "black box" nature of the model: interpretability can truly be a game changer in this setting. In addition to that, the new EU proposal for regulation on Artificial Intelligence (COM/2021/206) due to enter into force in the second half of 2022, requires algorithms that make decisions which "significantly affect" users, to provide explanation. As a result, interpretability will become a key requirement in high-risk applications such as diagnostic imaging. With this in mind, the remainder of the paper is organised as follows: Section 2 presents recent relevant work as well as research gaps; Section 3 identifies what we believe should be the key objectives the community should be focusing on; Section 4 outlines directions worth investigating, which we believe are promising in achieving the key objectives; Section 5 concludes by discussing opportunities and challenges ahead in the outlined vision for advances in neuro-symbolic approaches to human-centered explainability.

[1] Deep Learning Market: Focus on Medical Image Processing, 2020–2030, August 2020. Available at: https://www.rootsanalysis.com/reports/deep-learning-market.html.

2. State of the art

Motivated by rising concerns on the interpretability and accountability of deep learning systems in areas such as criminal justice [28] and diagnostics [53], in recent years neural-symbolic computing has become a very active topic of research focused on investigating the integration of learning from experience and reasoning about what has been learned [6].

The most promising directions of research in this area include: i) representing symbolic knowledge as a neural network using rule-based approaches inspired by inductive logic programming [10] and probabilistic databases [4] or by embedding first order logic symbols into tensors [42]; ii) learning to fine-tune symbolic rules based on the output of neural learning [33]; and iii) model-based integration of reasoning and learning which mainly focuses on propositional knowledge and forward reasoning [8,47]. The issue of explainability has only recently become pressing in neural-symbolic computing and researchers have started to look into knowledge extraction methods, but these approaches are mainly focused on specific layers of the network to reduce complexity [47], or approximation of knowledge distillation via soft-logic rules [20]. More recently, [34] looked at extracting concepts and symbols from clusters of CNN kernels to validate visual explanations with symbolic rules. Less consideration has been given to the extraction and validation of knowledge (including concepts, graphs, and rules) that can be used to understand the inner workings of a trained network (in other words, what is learned/encoded in deep representations), in order to explain in human terms what determined a given outcome, and in turn facilitate intervention. The success of transformers in Natural Language Processing has initiated a debate in the research community on attention mechanisms and their role in explainability [21,43,49]. Despite attention might help the interpretation of results, the risk of dismissing important complex relationships between features, concepts and outcomes makes it insufficient on its own for explainability purposes.

Network dissection is an interesting approach to abstract semantic concepts from a neural network, but it is mainly used to produce disentangled representations and it does not explain the correlation among those concepts with respect to a given class or specific input, nor does it indicate their contribution in decision making [2]. Model distillation approaches, in particular distillation into graphs, is a promising way of producing interpretable models of how a DNN operates [30]. Key approaches in this direction, however, focus on explaining the hierarchy of concepts [54] and do not use graph analysis to interpret the behaviour of the neural network. In our recent work we have proposed for the first time the characterisation of a co-activation graph which reflects the behaviour of a feed forward DNN, and we have used community analysis to relate neurons' activations with semantic similarity among output classes [19]. Building upon this work, we have investigated the use of link prediction on an extended version of the co-activation graph to generate local explanations in terms of semantic properties [18]. We are not yet there, but the outcome of our investigations convinced us that we need approaches which can go beyond simple model distillation and analysis, aiming at building a complete high-level symbolic abstraction of the network's inner working. Such a comprehensive representation can result from the combination of i) a knowledge graph distilled from the neural network (for global understanding), ii) external semantic knowledge about concepts and their relationships (for local understanding), and iii) logical rules that can cater for uncertainty, extracted (deductive) and learned (inductive) from neural representations (for approximation). This new rich and inherently interpretable representation would need to be validated against and reconciled with commonsense and domain specific knowledge (for robustness) before being used to generate explanations for humans. Once explanations are generated, it is important to design ways to collect and incorporate human feedback, and inject that back into the untrained neural model.

So far, the neural-symbolic community has focused on the design of tightly-coupled systems that can embed symbolic reasoning into neural learning. In our opinion and based on the taxonomy of neural-symbolic systems proposed by Henry Kautz in his talk titled "The third AI Summer" [25], more attention should be given to loosely coupled systems for a truly hybrid and explainable approach where learning high-dimensional probabilistic features via DNN happens in a continuous space and reasoning as well as qualitative representation of uncertainty via Knowledge Representation and Reasoning happens in a discrete space. Such an approach would reduce combinatorial complexity and go beyond local explanations enabling global understanding, human feedback, and control in a neuro-symbolic cycle [11]. The interplay between neural-symbolic systems and Graph Neural Network (GNN) has also been explored recently [27]. GNN could be a natural fit to combine cognitive and neural representations, but they cannot be considered capable of causal or deductive reasoning.

If we look at the use of DNNs in areas such as Medical Image Analysis, the majority of approaches to explainability are post-model and based on attribution of input features to output (producing partial explanations which are subject to human interpretation) or attention based, sometimes coupled with rule extraction (producing local explanations) [44]. Neither of these approaches provides the necessary level of trust for high-stake decision making such as medical diagnostics, as they do not provide human understandable insights on the model inner workings and how the model representation relates to prior knowledge. Recent interest in applying transformers in medical image analysis is documented in [14], but the survey indicates that transformers are not consistently better than CNN and being extremely computation-intensive and data-hungry, applying them where the availability of annotated data is scarce (often the case in medical imaging) can be a problem. Furthermore, only a few studies target explainability for visual transformers.

We believe the success of neuro-symbolic integration for explainability needs to rely on a comprehensive high-level symbolic representation of the DNN inner workings, validated with respect to domain and commonsense knowledge and integrated in a logical layer. Validation is paramount to assess the quality of explanations [11] and therefore it is not an aspect to be underestimated. Such validated symbolic representation is the one to be used to generate explanations: 1) for domain experts (causal and contextual) supporting their decision making (cognitive) process and allowing them to provide fine-grained feedback adjusting the logical layer, and 2) for AI experts, allowing them to calibrate the predictions via knowledge injection.

3. Objectives in neuro-symbolic approaches to human-centered explainability

Based on our characterisation of the research gaps in neuro-symbolic AI, we believe the ability to design approaches that can specifically tackle the need for explainability requires to aim at three key goals or main objectives as listed below:

- O1 Neural-cognitive mapping: reconcile low level image features with semantic concepts and relevant knowledge into a suitable representation space. Such representation should enable not only interpretation of low level features as semantic concepts, but also automated reasoning about their properties so as to enable validation and qualitative analysis of the deep representation, including characterisation of bias and errors. The proposed representation space should also support the characterisation of causality and compositionality as key enablers for the generation of explanations.
- O2 Neuro-Symbolic Extraction-Injection cycle and Explainability for AI experts: define a suitable knowledge injection mechanism to propagate newly discovered and validated semantic relations and constraints from the representation space in O1 across specific intermediate layers of the neural network during training, in order to make the representation learning process reproducible and explainable for AI experts.
- O3 Explainability for stakeholders/end users to close the Explanation-Feedback-Control cycle: use semantic interpretation of features, causality and compositionality from O1 to generate human-understandable explanations and leverage such explanations to collect specific human feedback in order to i) continuously adjust and enrich cognitive reasoning processes with data-driven insights, catering for uncertainty in the decision making process and ii) enabling human-control iterations in solving potential conflicts in the reconciled logical representation.

Note that O2 and O3 are both centered in the ability to unveil the behaviour of the neuro-symbolic system in performing a given classification task, but they support two complementary aspects of the neuro-symbolic cycle: objective O2 focuses on the neuro-symbolic extraction-injection cycle whereby explanatory rules and relations provided to AI experts should be tuned and injected into an untrained network to improve/adjust learning; objective O3 focuses on the Explanation-Feedback-Control cycle whereby explanations provided to end-users and domain experts should be validated with respect to background and common-sense knowledge and used to augment cognitive reasoning for decision support.

A graphical representation of these two aspects and the way they interact is illustrated in Fig. 1

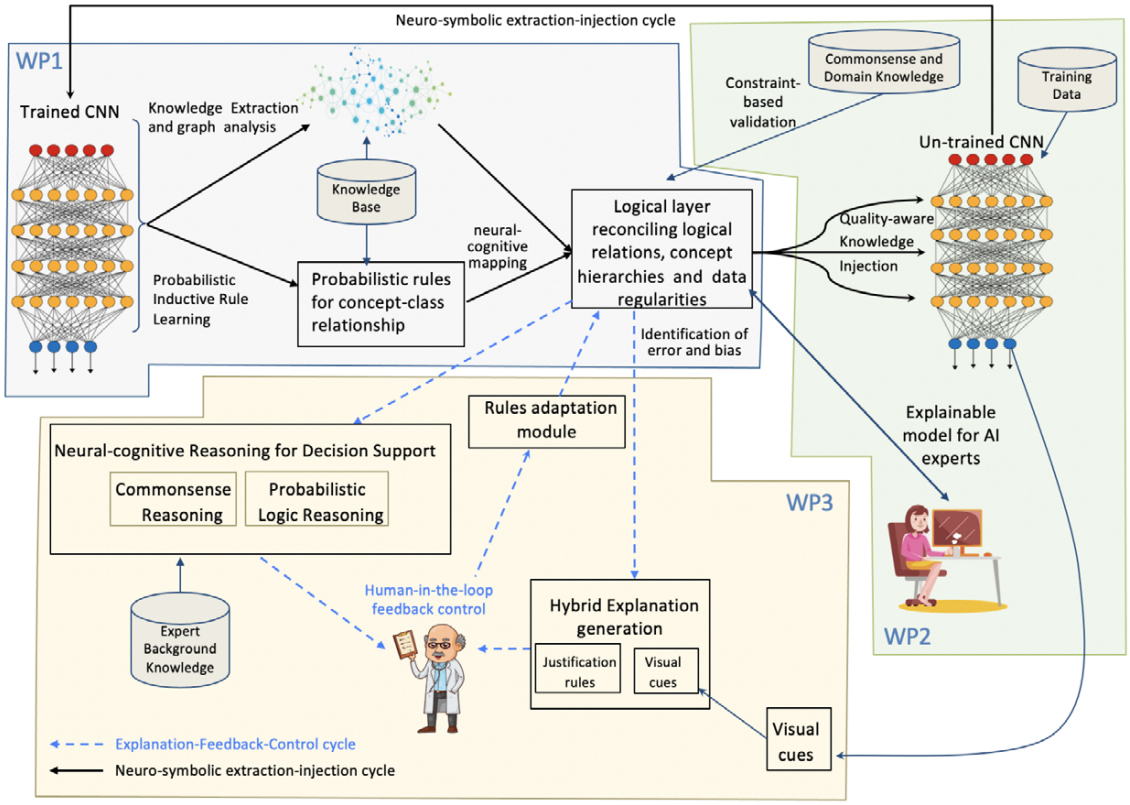

Fig. 1. Conceptual diagram illustrating the neuro-symbolic extraction-injection cycle and the complementary explanation-feedback-control cycle.

Our focus is determined by the consideration that the way humans learn is fundamentally different from the way we can teach machines to learn. New concepts in human processes are not learned in isolation, but considering connections to what is already known [15], and involving both cognitive and neural processes.

The research directions suggested in this paper focus on the design of solutions for combining neural learning and cognitive reasoning in a new hybrid model used to generate explanations considering cause-effect, compositionality and context [1] and leverage them to produce better models via explanatory interactive learning [46].

Such a neural-cognitive cycle exhibits the following key features:

- extraction and integration of knowledge from a trained network considers graphs, semantic concepts and probabilistic rules; these are extended, reconciled and validated with external relevant knowledge into a logic layer ensuring robustness and inherent interpretability;
- injection of knowledge not only in fully connected layers [13] but also in hidden layers aims at adjusting the way the entire network learns and reducing the impact of data-driven representation learning (including data bias and generalisation);
- explanation generation for domain experts goes beyond the combination of visual attention methods and question answering, as it exposes causality, hierarchies and logical relations learned by the deep model and reconciled in the logical layer;
- human-centered explainability leverages explanations to guide experts in providing targeted feedback to add/remove/modify conditions and rules in the logical layer enhancing interpretability and trust; this is expected to lead to wider adoption of Deep Learning for high-stake decision making, and contribute to materialising their potential in terms of efficient reporting and reduction of human errors.

The suggested directions for investigation in the development of new hybrid approaches to representation, learning and reasoning rely on the extraction of knowledge from a trained deep network to understand its inner workings

at a global level, the augmentation of such knowledge with commonsense relevant for a domain, and the injection of parts of the resulting augmented knowledge into the learning process for a non-trained network. Unlike existing solutions to explainability in Deep Learning which are only providing local model understanding and are therefore incomplete, these new approaches aim at considering the learning and decision process as a whole, through a holistic symbolic representation that makes it possible to go from one approach to the other and back without losing information in the conversion. A key step in the process is the involvement of domain experts (such as medical practitioners) as decision makers in validating and refining the intermediate representation based on the generated explainable outcome.

This new characterisation can drive fundamental changes and new opportunities in the next generation of algorithms and tools for machine intelligence from both a semantic and a big data perspective [1,3,22].

4. Key elements of the extraction-explanation-injection cycle

As a starting point in laying the foundations of our investigation, we outline three key assumptions or hypothesis (H) before defining the elements we believe should be part of the core of a neuro-symbolic cycle for human-centered explainability.

H1. *Graphs are well suited to extract knowledge about neural activities in deep networks, while probabilistic logic rules can intuitively and effectively represent causality, complex relational dependencies, and uncertainty: their combination is best suited as a formalism to reconcile commonsense priors and domain-knowledge with deep representations.*
Inspired by work done in neuroscience where functional graphs are used to represent the inner workings of the human brain, we have proposed a preliminary notion of co-activation graph in [19] that can represent statistical correlations among relevant features of any deep representation in a feed forward neural network. There is potential in this approach not only for representing co-activations, but also for analysing their relation with semantic concepts using, among others, community analysis and link prediction techniques. However, the complexity of relational and causal dependencies under uncertainty can be better captured by a formal language that combines first-order logic rules and statistical relational learning [40]. There are reasons to believe it is possible to construct a complex high-level symbolic representation that can reproduce some of the behaviour in the DNN and bridge the gap between human and AI [23].

H2. *If we can inject and propagate domain knowledge into any layer of a non-trained network during learning, such a hybrid approach to representation learning can produce inherently explainable models that are also robust and accurate.*
Vector embeddings as well as modification of the loss function have become popular approaches to inject background knowledge into DNNs with a focus on improving performance [5]. Since the structural and relational properties of the injected knowledge are lost once the knowledge is embedded or transformed, these approaches are not particularly helpful for explainability. To avoid this, the association between logical rules and semantic concepts on one hand, and semantic concepts and individual units on the other hand, need both to be explicit and accessible during the knowledge injection process: the former can rely on approaches such as inductive rule learning to adjust the weights of logical formulas by maximising the log likelihood of the training data, the latter can be done by leveraging approaches such as Network Dissection for disentangled representations.

H3. *The quality and trust in a hybrid neuro-symbolic cycle are enhanced by actively involving human experts in the process of semi-automatically and iteratively adjusting cognitive reasoning processes based on semantic interpretation of low level features and characterisation of their complex causal and relational dependencies.*
Assessment of quality and trust can be very difficult due to the subjective nature of the concept and lack of ground-truth. The quality and trust for an explanation has to be measured incrementally by domain experts: every time a control-feedback is requested to annotate or update specific rules in the reconciled representation (we refer to this process as *human-in-the-loop feedback control), variations of quality and trust have to be incrementally assessed through specific questionnaires and carefully designed assessment scores that are*

gender-specific. Focus group sessions with domain experts as well as engineers need to be held periodically to collect such feedback, considering accepted inaccuracies and subjectivity for image interpretation by domain experts (typically the ability of the model to accurately classify classes is in the order of 95–98%[2]). The aim should be to achieve comparable accuracies or surpass those achieved by human experts, at the same time increasing transparency and trust.

Based on the three key hypothesis above, the new class of approaches we suggest has to include three elements: neural-cognitive mapping, hybrid representation learning and explanation generation and feedback loop. In what follows we are going to describe what type of activities and approaches are suitable and can be considered for each of these elements.

4.1. Neural-cognitive mapping

When it comes to characterising the inner working of a trained deep network in a symbolic representation space, the combination of graph analysis and graph summarisation techniques with probabilistic rules extracted from deep representations can provide a complete and robust characterisation. One way to build and analyse the knowledge graph is to take inspiration from the use of functional graphs in neuroscience [12,32] building upon previous work in [18,19]. An interesting extension would be to explore the use of graph summarisation techniques to identify statistically relevant connections. The probabilistic rule extraction approach has to rely in some way on disentangled representation, such as those obtained by Network Dissection [2]. One possible way to achieve this would be to combine disentangled representation with some form of ranking of feature maps as in [9]. Concept occurrences can then be used as positive and negative examples for a probabilistic inductive rule learning system such as ILASP (Inductive Learning of Answer Set Programs) [29]. Results of graph analysis and summarisation then need to be reconciled with learned probabilistic logic rules, in order to characterise hierarchical relationships, causality and compositionality for the identification of errors and bias through human feedback. The resulting logical layer would be the main outcome of the neural-cognitive mapping process and it should encode the semantics of input, output and their causal and relational dependencies.

4.2. Hybrid representation learning

A representation learning mechanism referred to as *hybrid* would take as input the neural-cognitive mapping discussed in Section 4.1, validate it against commonsense and domain knowledge to assess the quality of the learned representation, and design a mechanism to inject (part of) such knowledge into the learning process of a non-trained network. In order to be effective, the layer-specific injection approach has to consider the concept hierarchy represented in the DNN's hidden layers: knowledge used at the highest layers is combined with task-specific feature representations (simpler but less generalisable), while knowledge used in the lower layers produces representations that are more complex but more general. According to that, semantic relationships (expressed as logical constraints) can be injected into different layers of the deep network. As mentioned earlier, association between constraints to be injected and layers has to leverage the alignment of hidden units with semantic concepts (disentanglement), and the relationships between those concepts in the constraints to be injected (rule induction). This association and the concept hierarchy have to be taken into account when devising a knowledge injection mechanism that can be based, for example, on the design of a *dom* heuristic (used in constraint satisfaction programming to minimise the depth of the search tree) [37] for Semantic Based Regularisation (SRB) [7], adapted for deep learning in Computer Vision. Learning can then be performed by minimising a semantic loss function [51] to maximise the log likelihood of the training data, so that the network is as close as possible to satisfying the semantic constraints at any given layer. Quality of the injected knowledge needs to be assessed by validating formal properties (such as consistency and completeness), and semantic properties (such as provenance) in relation to given domain-specific and commonsense knowledge. Based on the initial evaluation results, these metrics might require adjustment, which can result in best

[2]We refer to ROC AUC (Area Under the Receiver Operating Characteristic Curve), a metric used to assess the performance of classification models.

practices in data and knowledge curation as well as parameters setting for hybrid deep representation learning components.

4.3. Explanation generation, feedback and validation

The ability to close the neuro-symbolic cycle by generating explanations for domain experts and gather feedback is key for building as well as validating trustworthy and interpretable deep representations.

The neural-cognitive mapping aims to produce a consolidated symbolic representation space for the inner working of the trained deep network, and such representation can be leveraged to generate explanations for experts on certain specific outcomes as well as on the model as a whole. Such explanations, along with the same commonsense knowledge used to validate the consolidated representation and probabilistic inductive rules learned in the mapping process, guides the collection of human feedback.

In a similar approach, the neural-cognitive reasoning capability relies on the reconciled and validated representation and experts' background knowledge (if available), while the learning capability relies on experts' feedback used as new beliefs to adjust probabilistic logic rule weights [35,36]. The hybrid explanation generation capability does not have to solely rely on the symbolic representation, however, and can combine visual cues (such as saliency maps) from feature maps ranking, with causal explanations inspired by the notion of justification for logic programs [48] extended to probabilistic rules, and link prediction for Knowledge Graphs in order to generate textual factual and counterfactual explanations based on the outcome of neural-cognitive mapping. As a result, the approach will be able to address two gaps: considering causal-inference, and including domain experts in the process [45].

Validation of the effectiveness of the neuro-symbolic cycle for explainability is not a trivial task, and there is a pressing need for a benchmark for explanatory interactive learning [16]. These are known challenges in Explainable AI as discussed in Section 5. The identification of the most effective way to measure not only effectiveness but also level of trust in the model requires particular attention. It is worth highlighting that what we propose is a cyclic loop involving knowledge (extraction-injection cycle) as well as human feedback (explanation-feedback-control cycle). As such, we should consider comparing measures of quality and trust variations among subsequent iterations of the neuro-symbolic extraction-injection cycle as well as the Explanation-Feedback-Control cycle. One way of quantifying such measurements would be to adapt methods such as the System Causability Scale (SCS) [17] and Trustworthy Explainability Acceptance (TEA) [24].

The outcome of this human-centred feedback and adaptation process can be subject to bias, as perception can vary. The need to rely on a representative sample is important, as well as the need to mitigate bias with a process similar to inter-annotator agreement. Particular attention should be given to the gender dimension of bias. In this regard, it is important not only to make sure there is a representative sample of both genders in evaluators involved, but one should also consider using different types of questionnaires and scales according to the gender of the evaluator.

Careful design of such evaluation methodology, including consideration for gender-specific feedback, can result in best practices in explainable neural-cognitive learning and reasoning in specific fields such as clinical diagnostics.

5. Concluding remarks: Challenges and opportunities

This position paper discusses a new class of approaches for the design of a neuro-symbolic cycle that leverages symbolic knowledge, deductive reasoning and human feedback for explainability. The integration of cognitive and neural approaches to representation, learning and reasoning as discussed in this paper provides the scientific foundation for the creation of intrinsically explainable learning models. Explainability brings trust and accountability, creating new opportunities for Deep Learning to be widely applied to high-stake decision making such as clinical diagnostics. The remainder of this section summarises key opportunities opened up by the proposed research agenda as well as the open challenges.

5.1. Potential impact and opportunities

The ability to systematically and effectively address the issues of explainability and transparency of deep learning models has a huge potential in bridging the gap between advances in research and impact on the quality of life when high-risk decisions are involved.

If we consider diagnostic imaging as one of such critical application areas, there is evidence to support how domain experts do believe that the use of deep learning can actually speed up the processing and interpretation of radiology data by 20%, reducing the rate of false positives by approximately 10%. However, there is still a lack of clinical adoption due to opaqueness and lack of accountability of such models. The research directions outlined in this paper aim to address this gap, leading to increased trust, greater patient empowerment and, ultimately, better outcomes, including improved diagnosis, wider clinical deployment and greater efficiency in time and cost.

Beyond the evident potential impact on societies, health and well-being, there are also economic and commercial aspects to be considered. According to Cynthia Rudin [38] there is a fundamental problem with the business model of proprietary black-box deep learning systems: companies profiting from these proprietary models are not necessarily accountable for the quality of their results, therefore they are not incentivised to pay particular attention to the accuracy of their algorithms. There are several examples where the error of the model was not picked up, nor the developers made accountable for the effects of the model's error: recidivism risk prediction not considering the seriousness of the crime in COMPAS;[3] BreezoMeter's prediction of air quality as "ideal for outdoor activities" during the California wildfires of 2018 when it was dangerously bad;[4] incorrect diagnosis of pneumonia in chest radiographs as the model was picking up on the word 'portable' within the X-ray image, representing the type of X-ray equipment rather than the medical content of the image [53]. A truly explainable model is unlikely to suffer the risks of such an unnoticed catastrophic outcome. There is a broad target market for this change. If we look at diagnostic imaging only, a recent report[5] indicates that over 200 deep learning approaches are currently available for medical image processing, worth over 2 billion USD investments, which is projected to grow at a Compound Annual Growth Rate (CAGR) of 35% between 2020 and 2030.

In order to promote research innovation and faster adoption of this type of technology, implementations should be released Open Source either under MIT License or CC SA-BY-NC (Creative Commons share alike, attribution, non-commercial). New market opportunities can be generated by companies developing modifications of the Open Source version tailored to specific needs or added components, which can be licensed for specific tasks. The value for companies is that they will be able to sell a solution and be accountable for the quality of the model and its transparency, resulting in a better product, higher ratings and wider use. Longer term, the Open Source success will create new opportunities to change the business model of proprietary black-box deep learning systems, in that there will be other aspects customers will be paying for: not the software itself, but services associated with the use of the framework (e.g. training, consulting), or hosting and support (such as Cloud hosting and Open Source as-a-service). The latter is a particularly good fit given how computationally intensive modern deep learning approaches are.

5.2. Challenges ahead

Advances discussed in this paper towards human-centered explainability via neuro-symbolic AI do not come free of challenges, and the suggested directions of investigation presents some risks that are worth discussing, proposing ideas on how such risks could be mitigated.

The availability of good quality training datasets as well as the impact of knowledge quality could be problematic when trying to exploit the mapped symbolic representation for explanation generation. To address this issue some attention should be given to the design of specific quality metrics and the identification of suitable thresholds to reduce the impact of noisy knowledge and/or data. Human involvement not only in the design of such metrics but also on their quantitative and qualitative analysis can be of help at this stage of the process.

---

[3] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm – May 23, 2016.

[4] Mc Gough. M. (2018). How bad is Sacramento's air, exactly? Google results appear at odds with reality, some say. Sacramento Bee.

[5] https://www.rootsanalysis.com/reports/deep-learning-market.html

Table 1
Publicly available commonsense and domain knowledge datasets

Datasource	Category	Format	Size
Biobank UK	Multimodal (csv, txt, MRI, ...)	Multiple	500K participants (UK)
DOLCE	Upper Level Ontology	OWL/FOL	100 axioms (<20 MB)
DBPedia (EN)	Linked Open Dataset (multi-domain)	RDF	1.3Bi triples (34 GB)
OpenCyc	Commonsense Ontology	CycL / OWL	1.6M triples (~150 MB)
ConceptNet	Commonsense Knowledge Base	Graph	21M edges, >8M nodes(~20 GB)
EBI	Linked Open Dataset (domain-specific)	RDF	~6K triples
PubMed	Research Publications	Txt / XML	7.4M articles
Snomed CT	Clinical terminology system	RF2	353K concepts, 2.4M relations (~1.1 GB)

In terms of the symbolic representation suggested for neural-symbolic mapping, there are many different logical languages and semantics that can be considered: identifying the best candidate might not be a straightforward task. One rule of thumb should be to consider the level of expressivity vs. complexity required by the specific application domain and task, focusing on first-order logic languages and looking at the availability of efficient implementations of relevant reasoning engines.

The amount of data available to train the model might also present a challenge depending on the application scenario and the neural architecture used, as the most accurate of these models are known for being incredibly data-hungry. As an example, in the area of Cardiac MRI the ACDC dataset[6] represents a good starting point. In general, however, we argue that the ability to combine both knowledge and data represents an advantage of the suggested approaches as it makes them less likely to be affected by limited amounts of data or knowledge when one complements the other and appropriate validation is performed. It is good practice to consider publicly available datasets to be used as commonsense and clinical knowledge, which can serve as initial benchmarks of the methodology. These include semantic general knowledge such as DBPedia, Ontologies such as DOLCE,[7] OpenCyc[8] and ConceptNet,[9] textual data such as PubMed,[10] as well as domain-specific clinical knowledge such as Biobank UK,[11][12] EBI,[13] Snomed CT.[14] Table 1 summarises a few key useful characteristics of these datasets, including their content, format and size.

A final consideration needs to be made in relation to the challenge of engaging with the broader AI community in order to maximise the potential impact of solutions combining neural and symbolic approaches to AI. Recent efforts indicate the future looks promising and interdisciplinary research projects as well as initiatives bringing together experts with different AI background are flourishing. This gives us hope, but nonetheless diverging foundational phylosophies and background of the individual research groups involved should not be overlooked.

Acknowledgement

This research was conducted with the financial support of Science Foundation Ireland (12/RC/2289_P2) at Insight the SFI Research Centre for Data Analytics at Dublin City University.

[6] https://www.creatis.insa-lyon.fr/Challenge/acdc/
[7] http://www.loa.istc.cnr.it/dolce/overview.html
[8] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/dataset/opencyc
[9] http://conceptnet.io/
[10] https://www.nlm.nih.gov/databases/download/pubmed_medline.html
[11] https://www.ukbiobank.ac.uk/
[12] https://biobank.ctsu.ox.ac.uk/~bbdatan/Accessing_UKB_data_v2.3.pdf
[13] https://www.ebi.ac.uk/rdf/
[14] https://www.ehealthireland.ie/ehealth-functions/ehealth-standards-and-shared-care-records/standards-and-terminologies/snomed-ct/about-snomed/

References

[1] Z. Akata, D. Balliet, M. de Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, H. Hung, C. Jonker, C. Monz, M. Neerincx, F. Oliehoek, H. Prakken, S. Schlobach, L. van der Gaag, F. van Harmelen, H. van Hoof, B. van Riemsdijk, A. van Wynsberghe, R. Verbrugge, B. Verheij, P. Vossen and M. Welling, A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, *Adaptive, Responsible, and Explainable Artificial Intelligence, Computer* **53**(8) (2020), 18–28.

[2] D. Bau, B. Zhou, A. Khosla, A. Oliva and A. Torralba, Network dissection: Quantifying interpretability of deep visual representations, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3319–3327. doi:10.1109/CVPR.2017.354.

[3] A. Bernstein, J. Hendler and N. Noy, A new look at the Semantic Web, *Commun. ACM* **59**(9) (2016), 35–37. doi:10.1145/2890489.

[4] W.W. Cohen, F. Yang and K. Mazaitis, TensorLog: A probabilistic database implemented using deep-learning infrastructure, *J. Artif. Intell. Res.* **67** (2020), 285–325. doi:10.1613/jair.1.11944.

[5] T. Dash, S. Chitlangia, A. Ahuja and A. Srinivasan, A review of some techniques for inclusion of domain-knowledge into deep neural networks, *Scientific Reports* **12**(1) (2022), 1040. doi:10.1038/s41598-021-04590-0.

[6] A. d'Avila Garcez, M. Gori, L.C. Lamb, L. Serafini, M. Spranger and S.N. Tran, Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning, 2019.

[7] M. Diligenti, M. Gori and C. Saccà, Semantic-based regularization for learning and inference, *Artificial Intelligence* **244** (2017), 143–165. doi:10.1016/j.artint.2015.08.011.

[8] I. Donadello, L. Serafini and A. d'Avila Garcez, Logic tensor networks for semantic image interpretation, in: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 1596–1602. doi:10.24963/ijcai.2017/221.

[9] E. Ferreira dos Santos and A. Mileo, From disentangled representation to concept ranking: Interpreting deep representations in image classification tasks, in: *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Springer Nature, Switzerland, Cham, 2023, pp. 322–335. ISBN 978-3-031-23618-1. doi:10.1007/978-3-031-23618-1_22.

[10] M.V.M. França, G. Zaverucha and A. Garcez, Fast relational learning using bottom clause propositionalization with artificial neural networks, *Machine Learning* **94**(1) (2014), 81–104. doi:10.1007/s10994-013-5392-1.

[11] A.D. Garcez and L.C. Lamb, Neurosymbolic AI: The 3rd wave, *Artificial Intelligence Review* (2023). doi:10.1007/s10462-023-10448-w.

[12] J.O. Garcia, A. Ashourvan, S. Muldoon, J.M. Vettel and D.S. Bassett, Applications of community detection techniques to brain graphs: Algorithmic considerations and implications for neural function, *Proceedings of the IEEE* **106**(5) (2018), 846–867. doi:10.1109/JPROC.2017.2786710.

[13] E. Giunchiglia, M.C. Stoian and T. Lukasiewicz, Deep learning with logical constraints, in: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L.D. Raedt, ed., International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 5478–5485, Survey Track. doi:10.24963/ijcai.2022/767.

[14] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang and D. Shen, Transformers in medical image analysis, *Intelligent Medicine* **3**(1) (2023), 59–78. doi:10.1016/j.imed.2022.07.002.

[15] D. Hofstadter, *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*, 1995.

[16] L. Holmberg, *Towards Benchmarking Explainable Artificial Intelligence Methods*, 2022.

[17] A. Holzinger, A. Carrington and H. Müller, Measuring the quality of explanations: The System Causability Scale (SCS), *KI – Künstliche Intelligenz* **34**(2) (2020), 193–198. doi:10.1007/s13218-020-00636-z.

[18] V.A.C. Horta and A. Mileo, Generating local textual explanations for CNNs: A semantic approach based on knowledge graphs, in: *AIxIA 2021 - Advances in Artificial Intelligence - 20th International Conference of the Italian Association for Artificial Intelligence, December 1-3, 2021, Revised Selected Papers*, Lecture Notes in Computer Science **13196** (2021), 532–549. doi:10.1007/978-3-031-08421-8_37.

[19] V.A.C. Horta, I. Tiddi, S. Little and A. Mileo, Extracting knowledge from deep neural networks through graph analysis, *Future Generation Computer Systems* **120** (2021), 109–118. doi:10.1016/j.future.2021.02.009.

[20] Z. Hu, X. Ma, Z. Liu, E. Hovy and E. Xing, Harnessing deep neural networks with logic rules, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 2410–2420, https://aclanthology.org/P16-1228. doi:10.18653/v1/P16-1228.

[21] S. Jain and B.C. Wallace, Attention is not explanation, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3543–3556. (Long and Short Papers). doi:10.18653/v1/N19-1357.

[22] K. Janowicz, F. van Harmelen, J.A. Hendler and P. Hitzler, Why the data train needs semantic rails, *AI Magazine* **36**(1) (2015), 5–14, https://ojs.aaai.org/index.php/aimagazine/article/view/2560. doi:10.1609/aimag.v36i1.2560.

[23] 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, https://ojs.aaai.org/index.php/AAAI/article/view/21488. doi:10.1609/aaai.v36i11.21488.

[24] D. Kaur, S. Uslu, A. Durresi, S.V. Badve and M. Dundar, Trustworthy explainability acceptance: A new metric to measure the trustworthiness of interpretable AI medical diagnostic systems, in: *CISIS 2021*, Lecture Notes in Networks and Systems, Vol. 278, 2021.

[25] H.A. Kautz, The third AI summer: AAAI Robert S. Engelmore memorial lecture, *AI Mag.* **43**(1) (2022), 105–125. doi:10.1002/aaai.12036.

[26] M. Kim, J. Yun, Y. Cho, K. Shin, R. Jang, H. Bae and N. Kim, Deep learning in medical imaging, *Neurospine* **16**(4) (2019), 657–668. doi:10.14245/ns.1938396.198.

[27] L.C. Lamb, A.D. Garcez, M. Gori, M.O.R. Prates, P.H.C. Avelar and M.Y. Vardi, Graph neural networks meet neural-symbolic computing: A survey and perspective, in: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, C. Bessiere, ed., International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 4877–4884, Survey track. doi:10.24963/ijcai.2020/679.

[28] J.A. Larson, How We Analyzed the COMPAS Recidivism Algorithm, 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[29] M. Law, A. Russo and K. Broda, The ILASP system for Inductive Learning of Answer Set Programs, 2020.

[30] F. Lécué, On the role of knowledge graphs in explainable AI, *Semantic Web* **11**(1) (2020), 41–51, http://dblp.uni-trier.de/db/journals/semweb/semweb11.html#Lecue20. doi:10.3233/SW-190374.

[31] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, *nature* **521**(7553) (2015), 436. doi:10.1038/nature14539.

[32] J. Liu, M. Li, Y. Pan, W. Lan, R. Zheng, F.-X. Wu, J. Wang and M. De Domenico, Complex brain network analysis and its applications to brain disorders: A survey, *Complex* **2017** (2017). doi:10.1155/2017/8362741.

[33] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester and L.D. Raedt, DeepProbLog: Neural probabilistic logic programming, in: *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18*, Curran Associates Inc., Red Hook, NY, USA, 2018, pp. 3753–3763.

[34] K.H. Ngan, A.D. Garcez and J. Townsend, Extracting meaningful high-fidelity knowledge from convolutional neural networks, in: *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–17. doi:10.1109/IJCNN55064.2022.9892194.

[35] M. Nickles, diff-SAT – a Software for Sampling and Probabilistic Reasoning for SAT and Answer Set Programming, CoRR, 2021, arXiv:2101.00589.

[36] M. Nickles and A. Mileo, A system for probabilistic inductive answer set programming, in: *Scalable Uncertainty Management – 9th International Conference, SUM 2015*, Québec City, QC, Canada, September 16–18, 2015, C. Beierle and A. Dekhtyar, eds, Proceedings, Lecture Notes in Computer Science, Vol. 9310, Springer, 2015, pp. 99–105. doi:10.1007/978-3-319-23540-0_7.

[37] F. Rossi, P. van Beek and T. Walsh (eds), *Handbook of Constraint Programming, Foundations of Artificial Intelligence*, Vol. 2, Elsevier, 2006. ISBN 978-0-444-52726-4.

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

[39] Z. Salahuddin, H.C. Woodruff, A. Chatterjee and P. Lambin, Transparency of deep neural networks for medical image analysis: A review of interpretability methods, *Computers in Biology and Medicine* **140** (2022), 105111. doi:10.1016/j.compbiomed.2021.105111.

[40] A. Salam, R. Schwitter and M.A. Orgun, Probabilistic rule learning systems: A survey, *ACM Comput. Surv.* **54**(4) (2021), 79:1–79:16. doi:10.1145/3447581.

[41] T.J. Sejnowski, The unreasonable effectiveness of deep learning in artificial intelligence, *Proc. Natl. Acad. Sci. USA* **117**(48) (2020), 30033–30038. doi:10.1073/pnas.1907373117.

[42] L. Serafini, I. Donadello and A.S. d'Avila Garcez, Learning and reasoning in logic tensor networks: Theory and application to semantic image interpretation, in: *Proceedings of the Symposium on Applied Computing, SAC 2017*, Marrakech, Morocco, April 3–7, 2017, A. Seffah, B. Penzenstadler, C. Alves and X. Peng, eds, ACM, 2017, pp. 125–130. doi:10.1145/3019612.3019642.

[43] S. Serrano and N.A. Smith, Is attention interpretable? in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2931–2951, https://aclanthology.org/P19-1282. doi:10.18653/v1/P19-1282.

[44] A. Singh, S. Sengupta and V. Lakshminarayanan, Explainable deep learning models in medical image analysis, *J. Imaging* **6**(6) (2020), 52. doi:10.3390/jimaging6060052.

[45] I. Stepin, J.M. Alonso, A. Catalá and M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, *IEEE Access* **9** (2021), 11974–12001. doi:10.1109/ACCESS.2021.3051315.

[46] S. Teso, Ö. Alkan, W. Stammer and E. Daly, Leveraging explanations in interactive machine learning: An overview, *Frontiers in Artificial Intelligence* **6** (2023). doi:10.3389/frai.2023.1066049.

[47] S.N. Tran and A.S. d'Avila Garcez, Deep logic networks: Inserting and extracting knowledge from deep belief networks, *IEEE Trans. Neural Networks Learn. Syst.* **29**(2) (2018), 246–258. doi:10.1109/TNNLS.2016.2603784.

[48] C. Viegas Damásio, A. Analyti and G. Antoniou, Justifications for logic programming, in: *LPNMR 2013*, Springer-Verlag, Berlin, Heidelberg, 2013, pp. 530–542. ISBN 9783642405631. doi:10.1007/978-3-642-40564-8_53.

[49] S. Wiegreffe and Y. Pinter, Attention is not explanation, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 11–20, https://aclanthology.org/D19-1002. doi:10.18653/v1/D19-1002.

[50] N. Xie, G. Ras, M. van Gerven and D. Doran, Explainable Deep Learning: A Field Guide for the Uninitiated, CoRR, 2020, arXiv:2004.14545.

[51] J. Xu, Z. Zhang, T. Friedman, Y. Liang and G.V. den Broeck, A semantic loss function for deep learning with symbolic knowledge, in: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan*, Stockholm, Sweden, July 10–15, 2018, J.G. Dy and A. Krause, eds, Proceedings of Machine Learning Research, Vol. 80, PMLR, 2018, pp. 5498–5507.

[52] X. Yuan, P. He, Q. Zhu and X. Li, Adversarial examples: Attacks and defenses for deep learning, *IEEE Transactions on Neural Networks and Learning Systems* **30**(9) (2019), 2805–2824. doi:10.1109/TNNLS.2018.2886017.

[53] J.R. Zech, M.A. Badgeley, M. Liu, A.B. Costa, J.J. Titano and E.K. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study, *PLoS Medicine* **15** (2018). doi:10.1371/journal.pmed.1002683.

[54] Q. Zhang, R. Cao, F. Shi, Y.N. Wu and S. Zhu, Interpreting CNN knowledge via an explanatory graph, in: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, New Orleans, Louisiana, USA, February 2–7, 2018, S.A. McIlraith and K.Q. Weinberger, eds, AAAI Press, 2018, pp. 4454–4463, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17354.