Integrating and Reasoning with Data-Induced Information: Knowledge Bases of Axioms and Learned Models

Laura Papi[1]

[1]Sapienza University of Rome

Abstract

Machine Learning models are nowadays becoming increasingly widespread across a wide range of application domains, and the models that dominate the current scene mainly rely on purely sub-symbolic approaches. However, while effective, these models often face limitations and challenges that symbolic approaches could help overcome. These include the integration of formal logical reasoning to constrain and formally guarantee the behavior of the models, improve their interpretability and explainability, and ensure their trustworthiness. In this paper, we present a novel neuro-symbolic research path that bridges sub-symbolic Machine Learning classifiers with symbolic techniques from the Knowledge Representation and Reasoning field. Unlike previous approaches, our framework enables reasoning simultaneously on both the raw data features used by the classifiers and a symbolic intensional knowledge specified for the domain. This work describes the main technical results we have obtained, the ongoing advancements, and the future challenges we aim to address.

Keywords

Neuro-symbolic AI, Knowledge Representation and Reasoning

1. Introduction

The recent advances of deep learning approaches have led to remarkable success in a wide range of application domains, such as finance, healthcare, and law [1, 2]. However, despite their strong empirical performance, purely sub-symbolic models still present important limitations when it comes to employing them in critical domains, including limited interpretability, difficulty in guaranteeing compliance with formal constraints, and the inability to perform formal logical reasoning tasks [3, 4].

We argue that one key limitation lies in the available forms of interaction with these models, which are usually restricted. Specifically when it comes to Machine Learning (ML) classifier models, the majority of the approaches in the literature treat them as functions that assign a class to a specific instance. In this sense, the interaction with the model is *local*. However, these models, in order to perform the local classification task, are actually trained to learn a *global* knowledge over the domain. This global, inductively acquired knowledge is generally difficult to access, and we argue that techniques enabling interaction with it would be highly useful.

For this purpose, the approach we propose is based on a novel framework that combines the induced knowledge *(sub-symbolic)* of ML classifiers with a knowledge explicitly specified by logic-based formalisms over the domain *(symbolic)*. This framework is specifically designed to enable formal reasoning tasks over this combined form of knowledge, which frames this approach in the general category of Neuro-symbolic Artificial Intelligence (NSAI) [5] systems. At the core of this field is the goal of reconciling the sub-symbolic approaches typical of the ML scenario with the symbolic techniques of the Knowledge Representation and Reasoning field. Several different approaches have been proposed in this direction, combining the symbolic and sub-symbolic techniques to obtain hybrid systems that exploit the strengths of both while mitigating their weaknesses. These approaches are often organized in a taxonomy of several categories [6], but they can also be grouped in terms of two main perspectives. On one hand, some approaches integrate reasoning techniques directly with sub-symbolic models, which

Doctoral Consortium of the 23rd International Conference on Principles of Knowledge Representation and Reasoning (KR 2026 DC), July 20-23, 2026, Lisbon, Portugal

papi@diag.uniroma1.it (L. Papi)

0009-0003-2281-9500 (L. Papi)

allows them to enhance interpretability [7]. On the other hand, many approaches are post-hoc, applying reasoning only to the outputs of sub-symbolic models and treating the models as black boxes [8]. We argue that it would be useful to develop a comprehensive framework that allows performing both kinds of reasoning: on the symbolic level and over raw data features. This approach is still under-explored and worth investigating. Our proposed framework, described in the next sections, moves in this direction.

The framework we present in the following is grounded in the Ontology-Driven Knowledge Bases paradigm. The core principle of this paradigm is to represent knowledge via two complementary components: *intensional knowledge* and *extensional knowledge*. The intensional part is specified through an ontology and consists of a vocabulary of predicates together with a set of axioms over these predicates. These axioms model the semantic knowledge possessed over the domain of interest. The extensional part instead refers to the actual data populating the domain, specifying which objects are instances of which predicates, and has been represented in different forms across the literature [9, 10]. The intuition behind our research line is that sub-symbolic ML classifiers can be seen as high-level functions mapping raw data features to output labels, which can be associated with symbolic predicates. Under this perspective, a set of classifiers operating on the same domain of raw data features implicitly defines the extensional knowledge of a Knowledge Base, while an ontology defines its intensional part. Our proposed framework is oriented to enable formal reasoning tasks over this novel form of Ontology-Driven Knowledge Base.

The rest of this paper is structured as follows: Section 2 describes the theoretical framework we developed for this goal and presents the main technical results; Section 3 discusses preliminary experiments, which form a central part of our ongoing research; Finally, Section 4 summarizes the results obtained so far, outlines current progress, and highlights future directions and challenges.

2. Theoretical framework

In this section we present a framework that allows the integration of knowledge induced by ML classifiers with knowledge specified by logic-based formalisms, and enables reasoning tasks over this combined form of knowledge.

The framework is defined by first fixing a set $\mathbb{A}$ of attribute symbols. Each of these symbols $A_i \in \mathbb{A}$ is associated to an attribute domain D_{A_i} consisting of the non-empty set of possible values for that attribute (e.g. a finite set of *categories*, natural numbers $\mathbb{N}$, or reals $\mathbb{R}$). Each classifier of the framework operates on a tuple $\mathcal{A} = (A_1, ..., A_n)$ of attributes from $\mathbb{A}$, and we denote with $\mathbb{F}(\mathcal{A}) = D_{A_1} \times ... \times D_{A_n}$ the *feature space of* $\mathcal{A}$. Formally, the framework introduces the novel notion of Hybrid Knowledge Base (*HKB*). A *HKB* K is a pair $K = (O, \Psi)$, where O is an ontology and Ψ is a set of classifiers that operate on a tuple $\mathcal{A}$ of attributes from $\mathbb{A}$. For this framework, we are interested in ontologies that admit *atomic concepts*, i.e. unary predicates, and *atomic roles*, i.e. binary predicates. Hence, more formally, we say that Ψ contains: for each atomic concept C declared in O, a binary classifier of the form $\kappa_C : \mathbb{F}(\mathcal{A}) \to \{0, 1\}$; and for each atomic role R declared in O, a binary classifier over pairs of the form $\lambda_R : \mathbb{F}(\mathcal{A}) \times \mathbb{F}(\mathcal{A}) \to \{0, 1\}$. For the classification labels we assume the usual meaning that the label 1 (resp. 0) corresponds to the positive (resp. negative) class. Viewing this definition under the lens of Ontology-Driven Knowledge Bases, the classifiers, and consequently the feature space on which they operate, implicitly define the extensional part of the knowledge base.

Example 1. *Consider a medical scenario exploiting classifiers to determine whether a patient is diabetic (classifier κ_D), male (κ_M), or pregnant (κ_P). Also, assume the ontology O models the domain through a vocabulary containing the atomic concepts D, M, and P for diabetics, males, and pregnant patients, respectively. Furthermore, the intensional knowledge of O states that pregnant patients cannot be male ($P \sqsubseteq \neg M$). The HKB for this example is composed by O and the set of classifiers described above, where κ_C is assigned to the atomic concept C (for $C \in \{D, M, P\}$).*

It is immediate to notice that reasoning over multiple classifiers defined on the same domain can easily lead to inconsistencies.

Example 2. *Recall Example 1. There is no guarantee that a feature space element classified positively by κ_P will not also be classified positively by κ_M. Therefore, even in this simple scenario, the extensional knowledge induced by the classifiers can violate the intensional knowledge, i.e., inconsistencies may arise.*

In order to manage these inconsistencies we define specific semantics that make the reasoning possible. We define an *interpretation* for a *HKB* K as a pair $J = (I, f)$, where $I = (\Delta^I, \cdot^I)$ is a first-order interpretation for the predicates in O, and f is a function associating to each element of the feature space $\mathbb{F}(\mathcal{A})$ a (possibly empty) set of objects of the interpretation domain Δ^I. This notion of interpretation allows modeling the situation where a combination of feature values leads to violating the ontology. We model this case by associating that instance to an empty set of objects of the interpretation domain, and we say that the instance is *discarded by the interpretation*. Given an interpretation J, we denote the set of elements discarded by J with $\mathsf{disc}(J)$. On the other hand, associating a non-empty set of objects to an element of $\mathbb{F}(\mathcal{A})$, models the fact that one or more real world objects might have the same combination of values for the features in $\mathcal{A}$. Furthermore, we define a *model* for a *HKB* K, as an interpretation I such that it satisfies all the axioms of the ontology, and the next conditions hold: (i) for each concept C of the ontology, $C^I = \{o \mid \exists \bar{a}.\, \kappa_C(\bar{a}) = 1 \text{ and } o \in f(\bar{a})\}$; (ii) for each role R of the ontology, $R^I = \{(o_1, o_2) \mid \exists \bar{a}, \bar{b}.\, \lambda_R(\bar{a}, \bar{b}) = 1 \text{ and } o_1 \in f(\bar{a}) \text{ and } o_2 \in f(\bar{b})\}$. In other words, (i) and (ii) ensure that a model for a HKB is such that the extensions of the ontology predicates faithfully reflect the classifiers outputs.

Clearly, in order to make the reasoning task as informative as possible, our goal should be to retain as much information as possible. This automatically means trying to minimize the number of discarded elements, which leads us to the definition of *minimally-discarding models*. We say a model J for a *HKB* K is minimally-discarding if there is no model J' of K such that $\mathsf{disc}(J') \subset \mathsf{disc}(J)$.

However, it is not difficult to verify that the presence of role axioms in the ontology leads to the presence of multiple minimally-discarding models that discard different subsets of the feature space. Intuitively, when a role axiom, i.e., role inclusion or role disjointness, is violated, it is not necessary to discard both the instances participating in the role. In fact, discarding only one of them already gets rid of the pair in the extension of the role, and therefore, fixes the violation. However, if both the instances are not involved in any other violation, there is no preference for discarding one over the other, which naturally leads to having at least two different minimally-discarding models.

Example 3. *Recall Example 1, and assume to introduce two classifiers over pairs λ_{cD} and λ_{cB}, able to detect, respectively, whether pairs of patients are compatible blood donors or share a compatible blood type. Assume the ontology also asserts that pairs of patients that are compatible blood donors should also have a compatible blood type, i.e., $cD \sqsubseteq cB$. Now consider a pair of raw data vectors $(\bar{a}, \bar{b})$ classified positively by λ_{cD} but negatively by λ_{cB}. It is not difficult to verify that there exist at least two minimally-discarding models in this setting: one discarding $\bar{a}$ while retaining $\bar{b}$; and one doing the opposite, i.e., discarding $\bar{b}$ while retaining $\bar{a}$.*

In the presence of multiple minimally-discarding models, we studied two different semantics: a *cautious* one, that is reminiscent of classical skeptical reasoning over all repairs of a Knowledge Base [11], and considers only the answers obtainable by all minimally-discarding models; and one based on the WIDTIO (*When In Doubt Throw It Out*) approach, inspired by the IAR semantics for consistent query answering [12], which evaluates the query over the model defined as the intersection of all the minimally-discarding models. In the following we refer to the first one as *skeptical semantics*, and to the second one as *WIDTIO semantics*.

Example 4. *Recall Example 3, and assume an additional raw data vector $\bar{c}$ such that the pairs $(\bar{c}, \bar{a})$ and $(\bar{c}, \bar{b})$ are both classified positively by λ_{cD} and λ_{cB}. Now consider a query asking for all the patients that are eligible to donate blood to another patient, i.e., $q = \{(x) \mid \exists y.cD(x, y)\}$. Under the skeptical semantics, the answers to q would include the vector $\bar{c}$, since the model discarding $\bar{a}$ still retains the pair $(\bar{c}, \bar{b})$ in the extension of cD, while the model discarding $\bar{b}$ retains the pair $(\bar{c}, \bar{a})$. Under the WIDTIO semantics instead, the intersection of all the minimally-discarding models would by definition discard both $\bar{a}$ and $\bar{b}$, excluding $\bar{c}$ from the answers of the query q.*

2.1. Main Technical Results

Under both the presented semantics, we studied the complexity of the query answering task, focusing on the *classifiers complexity*, i.e. the complexity in which only the set of classifiers is regarded as the input, while the query and the ontology are fixed. Specifically, we assume *DL-Lite*$^{\neg}_{\text{RDFS}}$ ontologies [13], Multi-Layer Perceptrons classifiers, and queries in UCQ$^{\neq}$. The complexity results obtained are the following:

- *The non-trivial consistency problem is NP-complete.*
- *The skeptical entailment problem is* coNExpTime*-complete.*
- *The WIDTIO entailment problem is* Σ^p_2*-complete.*

In all cases the hardness already holds for queries in CQ. Interestingly, if we forbid the presence of roles in the ontology axioms, therefore considering a fragment of *DL-Lite*$^{\neg}_{\text{RDFS}}$ for the ontology, the skeptical entailment problem complexity drops down to NP-complete. A paper describing the full contribution of this formal framework definition was presented at AAAI 2026 and published in the proceedings of the conference [14].

3. Preliminary Experiments

We are currently developing an implementation of the presented framework, guided by the complexity results established in the theoretical analysis. These results naturally suggest different implementation strategies depending on the expressive power of the ontology. However, in all scenarios, classifiers are represented through their weight matrices, and encoded as arithmetic constraints into the solver.

Implementation via ASP From the computational complexity results for the WIDTIO entailment problem with *DL-Lite*$^{\neg}_{\text{RDFS}}$ ontologies, it follows that the problem can be encoded in Answer Set Programming (ASP). Therefore, the implementation of the framework in this setting can be carried out exploiting state-of-the-art ASP solvers such as *clingo* [15] or *DLV* [16].

We have developed a preliminary ASP encoding of the framework and set up some preliminary experiments. The results indicate that this approach is practically feasible for datasets of moderate size. As expected from the theoretical complexity, scalability is limited by the increasing of the feature space size, both in the number of attributes in the feature space and in the dimension of their domains. Nevertheless, the ASP-based implementation provides a faithful realization of the full framework and obtains reasonable computational times for datasets of moderate size.

Implementation via SAT When restricting the ontology to the fragment of *DL-Lite*$^{\neg}_{\text{RDFS}}$ that forbids the presence of roles in the axioms, the complexity of the WIDTIO entailment problem drops down to NP-complete, which makes the problem encodable as a boolean satisfiability problem, enabling an implementation based on SAT/SMT solvers such as *Z3* [17].

Preliminary experiments show that this approach allows to manage significantly larger datasets while keeping the computational times low. This confirms the practical advantage of restricting ontology expressivity when scalability is a priority.

Moreover, a fundamental observation arises: the entailment problem of a query over the defined framework can be seen as a form of neural network verification. Most neural network verification tools in the literature are developed to check the satisfiability of neural networks with respect to a given property. Our problem instead needs to verify whether a given query is satisfied by a set of networks and an ontology over the domain. Intuitively, the entailment problem can be reformulated as a verification task by constructing an appropriate network representation together with a corresponding set of logical properties.

This perspective enables the use of specialized neural network verification tools, such as *Marabou* [18], which are in fact SMT-based solvers, as *Z3* itself, but they are specifically optimized for common

non-linear activation functions such as ReLU. From some preliminary tests, it appears that encoding the problem with *Marabou* leads to significantly faster computational times compared to general-purpose SMT solvers. Furthermore, while *Z3* supports limited forms of non-linearity, which strongly limits the implementation, *Marabou* provides several commonly used activation functions, while keeping the computational times low.

To summarize the preliminary experiments section, both the ASP-based and the SAT/SMT-based implementations are currently under development. The preliminary results confirm the feasibility of the proposed framework and highlight promising directions for obtaining efficient implementations.

4. Future Work

In this work we introduced a formal framework that integrates knowledge induced by ML classifiers with knowledge specified by logic-based formalisms, and enables reasoning over this combined form of knowledge. We presented the computational complexity results of the query answering task under different semantics and ontology languages. Furthermore, guided by these complexity results, we set up some preliminary experiments that show promising directions for implementing the framework.

The first goal of our future work is to obtain a full implementation of the framework, exploring both the described settings: one with higher expressivity of the ontology language but limited scalability, and another one that, although limiting the ontology language, can handle more complex scenarios. Alongside the practical implementation of the framework, we are currently investigating whether our approach can be exploited to improve the solving of some well-established neuro-symbolic tasks. The intuition here is that the majority of NSAI tasks in the literature focus on solving problems over specific given input instances, while our framework provides an even higher expressive power. In fact, we are able to perform reasoning over the entire feature space of the classifiers, not only over a given input instance. We believe this could lead to solving even more complex tasks than the ones usually tackled in the NSAI literature.

Furthermore, another direction we are interested in investigating concerns the synthesis of new classifiers exploiting the existing classifiers and the intensional knowledge over the domain. Intuitively, instead of training new models from scratch, the framework allows the formulation of queries that combine existing classifiers, which could be used to define novel classification tasks without retraining and defining new models.

Finally, we plan to extend the formal framework to account for the probabilistic nature of ML outputs. The current formulation of the framework treats the outputs of the classifiers as strictly binary decisions, and we argue that it would be interesting to investigate how to adapt this framework to a probabilistic form of classification.

Overall, the long-term goal of this research is to explore a novel neuro-symbolic approach that allows reasoning at the same time over the sub-symbolic raw-data features on which the classifier operates, and over symbolic intensional knowledge of the domain. We argue this is an underexplored field in NSAI which could lead to promising and interesting results.

Declaration on Generative AI

During the preparation of this work, the author used GPT-4 in order to: Grammar and spelling check. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

Acknowledgments

The work presented in this paper was carried out under the supervision of Professor Marco Console and Professor Gianluca Cima, who are gratefully thanked for their help and support.

References

[1] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future, Stroke and Vascular Neurology 2 (2017).

[2] S. Bahoo, M. Cucculelli, X. Goga, J. Mondolo, Artificial intelligence in finance: a comprehensive review through bibliometric and content analysis, SN Business & Economics 4 (2023).

[3] C. Thames, Y. Sun, A survey of artificial intelligence approaches to safety and mission-critical systems, in: 2024 Integrated Communications, Navigation and Surveillance Conference (ICNS), 2024, pp. 1–12.

[4] European Parliament, Council of the European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), 2024. URL: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32024R1689.

[5] P. Hitzler, A. Eberhart, M. Ebrahimi, M. K. Sarker, L. Zhou, Neuro-symbolic approaches in artificial intelligence, National Science Review 9 (2022) nwac035. URL: https://doi.org/10.1093/nsr/nwac035. doi:10.1093/nsr/nwac035. arXiv:https://academic.oup.com/nsr/article-pdf/9/6/nwac035/43952953/nwac035.pdf.

[6] H. Kautz, The third ai summer: Aaai robert s. engelmore memorial lecture, AI Magazine 43 (2022) 105–125. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/19122. doi:10.1002/aaai.12036.

[7] M. Arenas, D. Baez, P. Barceló, J. Pérez, B. Subercaseaux, Foundations of symbolic languages for model interpretability, 2021. URL: https://arxiv.org/abs/2110.02376. arXiv:2110.02376.

[8] R. Manhaeve, S. Dumančić, A. Kimmig, T. Demeester, L. D. Raedt, Deepproblog: Neural probabilistic logic programming, 2018. URL: https://arxiv.org/abs/1805.10872. arXiv:1805.10872.

[9] M. Bienvenu, M. Ortiz, Ontology-mediated query answering with data-tractable description logics, in: Proceedings of the Eleventh International Summer School Tutorial Lectures (RW 2015), 2015, pp. 218–307.

[10] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, R. Rosati, Linking data to ontologies, in: S. Spaccapietra (Ed.), Journal on Data Semantics X, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 133–173.

[11] D. Lembo, M. Lenzerini, R. Rosati, M. Ruzzi, D. F. Savo, Inconsistency-tolerant semantics for description logics, in: Proceedings of the Fourth International Conference on Web Reasoning and Rule Systems (RR 2010), 2010, pp. 103–117.

[12] D. Lembo, M. Lenzerini, R. Rosati, M. Ruzzi, D. F. Savo, Inconsistency-tolerant query answering in ontology-based data access, Journal of Web Semantics 33 (2015) 3–29.

[13] G. Cima, M. Console, R. M. Delfino, M. Lenzerini, A. Poggi, Answering conjunctive queries with safe negation and inequalities over RDFS knowledge bases, in: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI 25), 2025, pp. 14824–14831.

[14] G. Cima, M. Console, L. Papi, Foundations of formal reasoning over knowledge bases combining symbolic and sub-symbolic knowledge, Proceedings of the AAAI Conference on Artificial Intelligence 40 (2026) 18994–19002. URL: https://ojs.aaai.org/index.php/AAAI/article/view/38971. doi:10.1609/aaai.v40i23.38971.

[15] M. Gebser, R. Kaminski, B. Kaufmann, T. Schaub, Multi-shot ASP solving with clingo, CoRR abs/1705.09811 (2017).

[16] W. T. Adrian, M. Alviano, F. Calimeri, B. Cuteri, C. Dodaro, W. Faber, D. Fuscà, N. Leone, M. Manna, S. Perri, F. Ricca, P. Veltri, J. Zangari, The asp system dlv: Advancements and applications, KI - Künstliche Intelligenz 32 (2018) 177–179. URL: https://api.semanticscholar.org/CorpusID:46886517.

[17] L. de Moura, N. Bjørner, Z3: an efficient smt solver, in: 2008 Tools and Algorithms for Construction and Analysis of Systems, Springer, Berlin, Heidelberg, 2008, pp. 337–340. URL: https://www.microsoft.com/en-us/research/publication/z3-an-efficient-smt-solver/.

[18] H. Wu, O. Isac, A. Zeljić, T. Tagomori, M. Daggitt, W. Kokke, I. Refaeli, G. Amir, K. Julian, S. Bassan, P. Huang, O. Lahav, M. Wu, M. Zhang, E. Komendantskaya, G. Katz, C. Barrett, Marabou 2.0: A versatile formal analyzer of neural networks, 2024. URL: https://arxiv.org/abs/2401.14461. arXiv:2401.14461.