Synergies Between Machine Learning and Reasoning - An Introduction by the Kay R. Amel group[1]

Ismaïl Baaj (baaj@cril.fr)[1]
Zied Bouraoui (bouraoui@cril.fr)[1]
Antoine Cornuéjols (antoine.cornuejols@agroparistech.fr)[2]
Thierry Denœux (thierry.denoeux@hds.utc.fr)[3]
Sébastien Destercke (sebastien.destercke@hds.utc.fr)[3]
Didier Dubois (dubois@irit.fr)[4]
Marie-Jeanne Lesot (Marie-Jeanne.Lesot@lip6.fr)[5]
João Marques-Silva (joao.marques-silva@univ-toulouse.fr)[4]
Jérôme Mengin (Jerome.Mengin@irit.fr)[4]
Henri Prade (prade@irit.fr)[4]
Steven Schockaert (schockaertS1@cardiff.ac.uk)[6]
Mathieu Serrurier (mathieu.serrurier@irit.fr)[4]
Olivier Strauss (strauss@lirmm.fr)[7]
Christel Vrain (Christel.Vrain@univ-orleans.fr)[8]

[1] Université d'Artois, UMR CNRS 8188, CRIL, Lens, France
[2] UMR MIA-Paris, AgroParisTech, INRA - UniversitéParis-Saclay, Paris, France
[3] Université de Technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France
[4] IRIT, CNRS, Université Paul Sabatier, Toulouse, France
[5] Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, France
[6] School of Computer Science & Informatics Cardiff University Cardiff, United Kingdom
[7] LIRMM Laboratory, UMR CNRS 5506, University of Montpellier II, France
[8] LIFO, Université d'Orléans, Orléans, France

April 29, 2024

Abstract

This paper proposes a tentative and original survey of meeting points between Knowledge Representation and Reasoning (KRR) and Machine Learning (ML), two areas which have been developed quite separately in the last four decades. First, some common concerns are identified and discussed such as the types of representation used, the roles of knowledge and data, the lack or the excess of information, or the need for explanations and causal understanding. Then, the survey is organised in seven sections covering most of the territory where KRR and ML meet. We start with a section dealing with prototypical approaches from the literature on learning and reasoning: Inductive Logic Programming, Statistical Relational Learning, and Neurosymbolic AI, where ideas from rule-based reasoning are combined with ML. Then we focus on the use of various forms of background knowledge in learning, ranging from additional regularisation terms in loss functions, to the problem of aligning symbolic and vector space representations, or the use of knowledge graphs for learning. Then, the next section describes how KRR notions may benefit to learning tasks. For instance, constraints can be used as in declarative data mining for influencing the learned patterns; or semantic features are exploited in low-shot learning to compensate for the lack of data; or yet we can take advantage of analogies for learning purposes. Conversely, another section investigates how ML methods may serve KRR goals. For instance, one may learn special kinds of rules such as default rules, fuzzy rules or threshold rules, or special types of information such as

[1] Kay R. Amel is the pen name of the working group "Apprentissage et Raisonnement" of the GDR ("Groupement De Recherche") named "Aspects Formels et Algorithmiques de l'Intelligence Artificielle", CNRS, France (https://www.gdria.fr/presentation/) now called GDR RADIA (for "Raisonnement, Apprentissage, et Décision en Intelligence Artificielle" - https://gdr-radia.cnrs.fr/). This paper is a fully revised, restructured and updated version of a collective report [86].

constraints, or preferences. The section also covers formal concept analysis and rough sets-based methods. Yet another section reviews various interactions between Automated Reasoning and ML, such as the use of ML methods in SAT solving to make reasoning faster. Then a section deals with works related to model accountability, including explainability and interpretability, fairness and robustness. Finally, a section covers works on handling imperfect or incomplete data, including the problem of learning from uncertain or coarse data, the use of belief functions for regression, a revision-based view of the EM algorithm, the use of possibility theory in statistics, or the learning of imprecise models. This paper thus aims at a better mutual understanding of research in KRR and ML, and how they can cooperate. The paper is completed by an abundant bibliography.

1 Introduction

Learning and reasoning are two basic aspects of intelligence. In the context of Artificial Intelligence (AI), these two aspects have often been studied independently, giving rise to distinct fields of research: Machine Learning (ML) and Knowledge Representation and Reasoning (KRR), respectively. Despite the traditional separation between these two fields, there is now a welcome and growing emphasis in the literature on the complementary strengths and weaknesses of their respective methodologies. For instance, ML methods can deal with raw data (e.g., in textual or visual form) and often requires less modelling efforts to be deployed, as long as sufficient training data can readily be obtained, but they often lack interpretability, and, typically, cannot provide strong guarantees as to the robustness of their outputs. This black box nature of ML methods is a fundamental concern, which reduces their appeal in cases where high-stakes decisions have to be made [536]. The presence of gender and racial biases in data, for instance, is problematic whenever ML methods are deployed [584]. On the other hand, KRR methods tend to produce results in a more transparent and systematic way, but they typically rely on the availability of structured knowledge that has been carefully encoded in some formal language, and may face scalability issues. The so-called knowledge acquisition bottleneck is again a fundamental concern, which has played a central role in the more limited acceptance of KRR methodologies in the industry.

The need to integrate methods from ML and KRR, given their complementary nature, has already been extensively discussed in survey papers [64], edited volumes [310], special issues[1], a recurring special session at the KR conference[2], as well as dedicated workshops and conferences [156]. In this overview paper, we aim to contribute to this discussion by taking a broader view on the possible synergies. We note, in particular, that existing work tends to rely on generalised and overly simplistic dichotomies, which suggest that there exists a large gap between KRR and ML: KRR deals with knowledge, ML handles data; KRR privileges symbolic approaches, while numerical methods dominate in ML. Even if such claims cannot be fully denied, they are nonetheless misleading. We believe that they stem from an overly narrow view of what constitutes reasoning, which is often implicitly equated with the use of rule-based methods, and learning, which is increasingly being equated with neural-network-based methods. To illustrate this point, let us take the example of Case-Based Reasoning [2], where we need to make a prediction about some query case by relying on similar cases with known labels. Here, the reasoning aspect essentially involves a form of analogical transfer, i.e. inferring how the labels of similar cases need to be adapted given their differences with the query case. Here we are thus reasoning about data, rather than about structured knowledge, which may moreover be numerical. Such examples illustrate the claim that the boundary between learning and reasoning is blurrier than is often assumed.

While we will not attempt to precisely define what reasoning means, we note that it involves dealing with incomplete knowledge. In the traditional setting, a symbolic knowledge base encodes incomplete knowledge about a state of affairs in the form of a set of possible worlds. Similarly, in case-based reasoning we have highly incomplete knowledge of how labels have been assigned to cases, merely relying on the knowledge that the label assignments satisfy some kind of regularity which allows for analogical transfer. Of course, ML settings also involve incomplete knowledge and data, and learnt models represent imperfect knowledge about the “true” underlying function or distribution.

Perhaps, the dichotomy between KRR and ML somewhat echoes the distinction advocated in psychology by Kahneman between “System 1” and “System 2” for describing the two kinds of activities of the

[1] https://www.springer.com/journal/10994/updates/17562232

[2] https://kr2022.cs.tu-dortmund.de/cfp_special_session_kr_and_machine_learning.php

human mind: “*System 1 operates automatically and quickly, with little or no effort and no sense of voluntary control. System 2 allocates attention to the effortful mental activities that demand it, including complex computations. The operations of System 2 are often associated with the subjective experience of agency, choice, and concentration.*" [358]. Still, the relevance of this parallel is a matter of debate.

The aim of the present survey is to illustrate that synergies between KRR and ML are much broader and more diverse than commonly thought and go beyond the issues discussed in the previous paragraphs. We start in Section 2 with a presentation of a number of issues that arise both in KRR and ML. These common concerns include the importance of representations, the need for trade-offs to manage complexity, the need to deal with imperfect information, and the importance of explanations. Then, our survey is organised in seven sections as follows:

- In Section 3, we first focus on methods where ML is combined with rule-based reasoning, which covers the most prototypical approaches from the literature on Learning and Reasoning: Inductive Logic Programming (ILP), which is a family of machine learning methods where learned models take the form of a logic program, Statistical Relational Learning, which builds on ILP by combining logical representations and probability distributions for representing learned models, and Neurosymbolic AI, where ideas from rule-based reasoning are combined with neural networks.

- In Section 4, we then focus on cases where symbolic background knowledge, in some form, is used to improve traditional machine learning systems. This covers, for instance, cases where the available background knowledge gives rise to an additional regularisation term in the loss function, or where such knowledge is used to constrain the latent representations which are learned by the model, by making them more semantically meaningful (or even interpretable) in some sense.

- Section 5 subsequently discusses machine learning methods which rely on reasoning processes, beyond rule-based reasoning. This section covers work on declarative data mining, where some form of constraint satisfaction is used to influence or restrict the patterns which are learned (e.g. clusters) and low-shot learning, where reasoning is used to alleviate the lack of explicit training data. Under the same umbrella, we also consider case-based reasoning, analogical reasoning and transfer learning.

- Whereas the preceding sections are essentially concerned with using methods and insights from KRR to improve ML systems, the next two sections are mostly concerned with the use of ML methods for improving KRR systems. In Section 6, we discuss in particular how machine learning methods can be used to alleviate the knowledge acquisition bottleneck. The idea of learning rules from data plays a central role here. Rather than giving a comprehensive overview on rule learning, however, we emphasise in particular the need to go beyond traditional rules (e.g., Horn rules interpreted using material implication). We discuss in particular the advantages of learning default rules, fuzzy rules and threshold rules. Beyond rules, we also cover methods for learning constraints and preferences.

- Section 7 then surveys work in which ML methods are used for improving reasoning processes themselves. This covers traditional KRR settings, such as SAT solving, where ML methods may be used to make reasoning faster.

- In Section 8, we cover work related to model accountability, including explainability and interpretability, fairness and robustness, since the solution to such issues often relies on ideas from KRR. For instance, the need for interpretability may require us to relate learned models to symbolic knowledge, while robustness is linked to notions such as causality.

- Finally, Section 9 covers works on handling imperfect data including, for instance, the problem of learning from coarse labels. While this topic is not usually considered in the context of Learning & Reasoning, it is nonetheless highly relevant if we view reasoning from the perspective of dealing with incompleteness.

2 Common Concerns

In order to suggest and illustrate differences and also similarities between KRR and ML, let us start with the simple example of a classification or recommendation-like task, such as, e.g., associating the profile of a candidate (in terms of skills, tastes, and so on) with possible activities suitable for him/her in a vocational guidance system. Such a problem may be envisioned in different manners. On the one hand, one may think of it in terms of a rule-based system relying on some expertise (where rules may be pervaded with uncertainty), or on the other hand in terms of machine learning by exploiting a collection of data (here pertaining to past cases in career guidance).

Let us first observe that traditionally KRR is seen as mostly concerned by deductive and abductive reasoning while ML is mostly concerned by inductive reasoning. Beyond the differences of types of representation and reasoning that are used in both kinds of approach (e.g., conditional tables for uncertainty assessment vs. weights in a neural net), there are some noticeable similarities between (graphical) structures that can be associated with a rule-based reasoning device, handling uncertainty (or an information fusion process) and with a neural net. This remark suggests that, beyond differences in perspective, there is some structural resemblance between the two types of process. This resemblance has been investigated recently in detail in the setting of belief function theory [179], but an example may also be found in possibility theory, starting with an older work on a possibilistic (max-min) matrix calculus devoted to explainability (where each matrix represents a rule) [239, 213, 37].

Beyond this kind of parallel, KRR and ML have common concerns. This section gives an overview of the main ones regarding the representation issues, the complexity, the role of knowledge, the handling of lack of information, or information in excess, uncertainty, and last but not least regarding causality and explanation. Each subsection below tries to follow the same basic structure, by each time providing i) the KRR view, ii) the ML view, and iii) some synthesis and discussion.

2.1 Types of Representation

AI deals with studying and designing computer programs that behave intelligently, which in part entails mimicking mental faculties observed in living organisms like thinking and learning. Within AI, KRR is concerned with how knowledge can be represented symbolically and manipulated in an automated way by programs while, in ML, the focus is on discovering means, whatever they may be, for realizing induction.

In KRR, the central concerns are what an agent need to know to behave intelligently and what sort of computational mechanisms might allow the useful pieces of knowledge to be made available when required and be manipulated in order to produce new knowledge. In one word, reasoning is the aim.

In ML, symmetrically, the question is how to acquire knowledge from data in order to automatically produce what may be called programs that output useful and reasonable answers when fed with inputs or queries (e.g. propose a diagnosis when seeing a mammography from a patient). Here, learning is the keyword.

In both cases, knowledge plays a role, and an important one. But this is not the same one. It might thus appear that on one side are gardeners who meticulously tend to and extend “jardins à la française” while on the other side are barbarians for whom anything goes as long as seeds grow and who have no respect for the rules of plant growing and marrying. However, this is only a misconception: hard rules are followed and obeyed on each side.

On **the side of KRR**, the crux is the interplay between representation and reasoning, that is how knowledge can be represented as comprehensively as possible and, at the same time, be reasoned with as effectively as possible. There exists a whole range of representation systems and formalisms aimed at various uses and applications. Without exhaustiveness, one can mention logic-based ones like propositional logic, first-order logic, Horn clauses, default logic, fuzzy logic and its variants, or production rules, object-oriented and graph-based ones, and situation calculus. It can therefore appear that each of these representation systems is dedicated to a particular view on the world and a particular use. But then, why not look for a more general language, one that would allow to represent anything in the world and be, in a way, co-extensive with the natural language that we use to communicate between us? Brachman and Levesque [403, 402, 401] have shown that there is a trade-off between the expressiveness of a representation system and the computational tractability of reasoning using it.

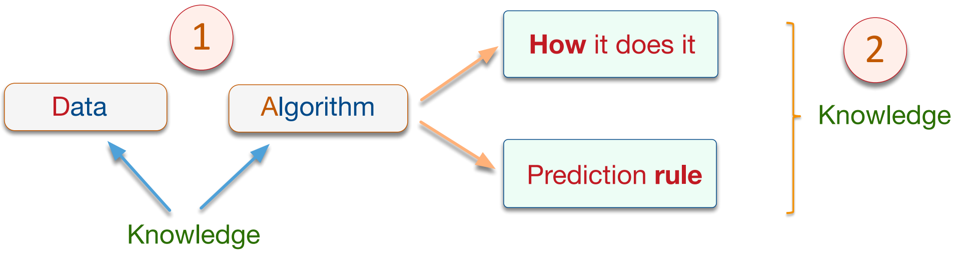

Figure 1: Machine learning transforms knowledge provided as input (1) under the form of data and the choice of a learning algorithm and produces knowledge (2) in the form, e.g., of prediction rules and possibly of information about the learning process followed.

There is thus a need for limiting reasoning to preserve tractability. For instance, not looking for all natural consequences of what is expressed using weaker notions of entailment.

On **the side of ML**, the tradeoff stems from a fundamental law: machine learning just reformulates what has been given as inputs. A kind of conservation theorem is at play: no information is "added" above the data provided and the existing prior knowledge. If, therefore, the data is limited, lots of prior knowledge is required, whereas big data allows for less prior knowledge. Machine learning "consumes" knowledge as input in the form of data, algorithms, which is a kind of knowledge, and prior knowledge and it produces knowledge in the form of prediction rules or patterns and possibly in the form of justification or explanation of the process used in this production (see Figure 1).

Here then, in order to control the quality of the induction that is realized by the machine, the ML practitioner must pay attention to the way the (never) raw data is fed which includes all the bias introduced by the choice of experimental apparatus, the choice of descriptors, their normalization, the possible enrichments using ontologies, the way missing values are dealt with, and so on, as well as the knowledge embedded in the learning algorithm and its own bias. This includes constraints on the hypothesis space, like the preference for sparse linear rules, as well as the choice of the architecture of the neural network if such a machinery is used.

ML has, therefore, its own tradeoff to delicately optimize between the information to provide in the form of data and the knowledge given by the choice of the learning algorithm. And because it is focussed on induction, an inference mechanism notoriously prone to errors, ML must be very careful about the quality of the results of learning, hence the emphasis on interpretability and explanations.

Looking back at representational issues, in KRR, as suggested by the name, the main representation issues concern the representation of pieces of knowledge (rather than data). The large variety of real world information has led to a number of logical formalisms ranging from classical logic (especially propositional and first order) to modal logics (for dealing with, e.g., time, deontic, or epistemic notions) and to non-classical logics for handling commonsense reasoning.

The representation may use different formats, directed or undirected: sets of if-then rules, or sets of logical formulas. A rule "if A then B" is a 3-valued entity (as first noticed in [164]), since it induces a partition between its set of examples, its set of counterexamples and the set of items for which the rule is irrelevant (i.e., when A is false). So a rule strongly departs from its apparent logical counterpart in terms of material implication $A \to B$ (which is indeed non-directed, since it is equivalent to $\neg B \to \neg A$). This discrepancy can be also observed in the probabilistic setting, since $Prob(B|A) \neq Prob(A \to B)$ in general. Rules may hold up to (implicit) exceptions (see subsection 2.3).

Knowledge may be pervaded with uncertainty, which can be handled in different settings, in terms of probability, possibility, belief functions, or imprecise probabilities (see subsection 2.3). In all of these cases, a joint distribution can be decomposed in sub-distributions laying bare some form of conditional independence relations, with a graphical counterpart; the prototypical graphical models in each representation are respectively Bayesian networks (probabilistic), possibilistic networks, credal networks (imprecise probabilities [139]) or valuation-based systems (belief functions [570]). Conceptual graphs [583, 111] offer a graph representation for logic, especially for ontologies/description logics.

The main goal of KRR is to develop sound and (as far as possible) complete inference mechanisms to draw conclusions from generic knowledge and factual data, in a given representation setting [299, 298].

The mathematical tools underlying KRR are those of logic and uncertainty theories, and more generally discrete mathematics. An important issue in KRR is to find good compromises between the expressiveness of the representation and the computational tractability for inferring the conclusions of interest from it [403]. This concern is especially at work with description logics that are bound to use tractable fragments of first order logic. (See subsection 2.2.)

The situation in ML is quite different concerning representation issues. ML aims at learning a model of the world from data. There are thus two key representation problems: the representation of data and the representation of models. See, e.g., [127, 128]. In many approaches the data space is assimilated to a subset of $\mathbb{R}^p$, in which the observations are described by p numerical attributes. This is the simplest case, allowing the use of mathematical results in linear algebra and in continuous optimization. Nevertheless, data may also be described by qualitative attributes, as for instance binary attributes, thus either requiring to transform discrete attributes into continuous numerical scales, or requiring different mathematical approaches, based on discrete optimisation and on enumeration coupled with efficient pruning strategies. Quite often, data is described by both types of attributes and only few ML tools, for instance decision trees, are able to handle them without any transformation into one single type. Therefore, changes of representation are needed, as for instance discretization, or the encoding of qualitative attributes into numerical ones, all inducing a bias on the learning process. More complex data, such as relational data, trees, and graphs need more powerful representation languages, such as first order logic or some proper representation trick as for instance propositionalization or the definition of appropriate kernels. It is important to notice that the more sophisticated the representation language, the more complex the inference process and a trade-off must be found between the granularity of the representation and the efficiency of the ML tool.

Regarding models, they depend on the ML task: supervised or unsupervised classification, reinforcement learning, learning to rank, mining frequent patterns, etc. They depend also on the type of approach that one favours: more statistically or more artificial-intelligence oriented. There is usually a distinction between *generative* and *discriminative* models (or decision functions). In the *generative approach*, one tries to learn a probability distribution $\mathbf{p}_{\mathscr{X}}$ over the input space $\mathscr{X}$, or at least a model able to generate samples assumed to follow a distribution close enough to $\mathbf{p}_{\mathscr{X}}$. If learning a precise and accurate enough probability distribution is successful, it becomes possible in principle to generate further examples $\mathbf{x} \in \mathscr{X}$, the distribution of which is indistinguishable from the true underlying distribution. It is sometimes claimed that this capability makes the generative approach "explicative", yet this is clearly a matter of debate. The *discriminative* approach does not try to learn a model that allows the generation of more examples. It only provides either a means of deciding, when in the supervised mode, or a means to express some regularities in the data set in the unsupervised mode. These regularities, as well as these decision functions can be expressed in terms of logical rules, graphs, neural networks, etc. While they do not allow to generate new examples, they nonetheless can be much more interpretable than probability distributions.

Very sketchily, one can distinguish between the following types of representations.

- Linear models and their generalisations, such as linear regression or the linear perceptron first proposed by Rosenblatt [531]. Because these models are based on linear weightings of the descriptors of the entries, it looks easy to estimate the importance of each descriptor and thus to offer some understanding of the phenomenon at hand. This, however, assumes that the descriptors are uncorrelated and are well chosen.

- Nonlinear models are often necessary in order to account for the intricacies of the world. Neural networks, nowadays involving very numerous layers of non linearity, are presently the favourite tools for representing and learning non linear models.

- Linear models as well as nonlinear ones provide a description of the world or of decision rules through (finite) combinations of descriptors. They are parametric models. Another approach is to approximate the world by learning a non previously fixed number of prototypes and use a nearest-neighbour technique to define decision functions. These systems are capable of handling any number of prototypes as long as they can fit the data appropriately. *Support Vector Machines* (SVM) fall in this category since they adjust the number of support vectors (learning examples) in order to fit the data. Here, explaining a rule may mean providing a list of the most relevant prototypes that the rule uses.

- The above models are generally numerical in essence, and the associated learning mechanisms most often rely on some optimisation process over the space of parameters. Another class of models relies on logical descriptions, e.g., sets of clauses. Decision trees can also be considered as logic-based, since each tree can be transformed into a set of clauses. The learning algorithms use more powerful structures over the space of models than numerical models. In many cases the discrete nature of the search space and the definition of a generality relation between formulas allow the organization of models in a lattice and the design of heuristics to efficiently prune the search space. More generally, these approaches are usually modeled as enumeration problems (e.g., pattern mining) or discrete optimization problems (supervised learning, clustering). Moreover such models offer more opportunities to influence the learning process using prior knowledge. Finally, they can be easily interpreted. The downside is their increased brittleness when coping with noisy data.

2.2 Coping with complexity

In both ML and KRR there is a trade-off between the generality of the approach and what is feasible in practice. In KRR, this generality amounts to the expressive power of the representation language which is used, while what is feasible is related to the computational complexity. In ML, this is related to the richness of the hypothesis space, and what is feasible is related to the amount of data that is available.

Complexity issues are a major concern in any branch of computer science. In KRR, very expressive representation languages have been studied, but interesting reasoning problems for these languages are often at least at the second level of the polynomial hierarchy for time complexity. There is a trade-off between the expressive power of a language and the complexity of the inference it allows. Reasoning tasks in languages with suitably restricted expressiveness are tractable, like for instance languages using Horn clauses or Lightweight description logics such as DL-lite [99] or EL [32].

The study of complexity has motivated a large number of works in many fields of KRR including non-monotonic reasoning, argumentation, belief merging and uncertainty management. In particular when the desirable solution (i.e., gold standard) of the problem (for instance, merging operator, inconsistency-tolerant consequence relation, etc.) has a high computational complexity, then it is common to look for an approximation that has reasonable complexity. For instance, the observation that answering meaningful queries from an inconsistent DL-Lite knowledge base using universal consequence relation is NP-Complete, has led to the introduction of several tractable approximations [41].

The attempt to cope with hardness of inference has also been a driving force in research around some important and expressive languages, including propositional clauses and CSPs, where inference is NP-complete; for instance, powerful methods nowadays enable the solving of SAT problems with up to hundreds of thousands of variables, and millions of clauses in a few minutes (see section 7.1). Some of the most competitive current SAT solvers are described in [4, 446, 427]. Two other ways to cope with time complexity are anytime methods, which can be interrupted at any time during the solving process and then return an incomplete, possibly false or sub-optimal solution; and approximate methods. A recent trend in KRR is to study so-called *compilation schemes* [149, 442]: the idea here is to pre-process some pieces of the available information in order to improve the computational efficiency (especially, the time complexity) of some tasks; this pre-processing leads to a representation in a language where reasoning tasks can be performed in polynomial time (at the cost of a theoretical blow up in worst-case space complexity, which fortunately does not often happen in practice).

Contrastingly, ML algorithms often have a time complexity which is polynomial in the number of variables, the size of the dataset and the size of the model being learnt, especially when the domains are continuous. However, because of the possible huge size of the dataset or of the models, capping the degree of the polynomial remains an important issue. In the case of discrete domains, finding the optimal model, i.e., the one that best fits a given set of examples, can be hard (see [338]), but one is often happy with finding a "good enough" model in polynomial time: there is no absolute guarantee that the model that best fits the training examples is the theoretical best model anyway, since this may depend on the set of examples. In fact, an important aspect of complexity in ML concerns the prediction of the quality of the model that one can learn from a given dataset: in the PAC setting for instance [610], one tries to estimate how many examples are needed to guarantee that the model learnt will be, with a high probability, a close approximation to the unknown target model. Intuitively, the more expressive the hypothesis space is, the more difficult it will be to correctly identify the target model, and the more examples will be needed for that [612], but the more likely it is that the right hypothesis is within the chosen space.

2.3 Imperfect information

The imperfection of information can take two opposite forms: information may be incomplete or insufficient; or it may be in excess, hence, wrong or conflicting. In the case of data, there may be missing values, or scarcity of observations, or on the contrary the data may be noisy or containing outliers. For knowledge bases, incompleteness rather manifests itself by the impossibility to infer a conclusion or its negation. Conflicting knowledge may be due to disagreeing sources of information, requiring suitable inconsistency management tools. Both situations (information missing or in excess) possibly generate uncertainty.

Uncertainty has always been an important topic in KRR [492][298]. While in ML uncertainty is almost always considered to be of statistical or probabilistic origin (often termed "aleatory uncertainty"), other causes for uncertainty exist, such as the sheer lack of knowledge, and the excess of information leading to conflicts (often termed "epistemic uncertainty"). However, the role of uncertainty handling in KRR and in ML seems to have been very different so far. While it has been an important issue in KRR and has generated a lot of novel contributions beyond classical logic and probability, it has been considered almost only from a purely statistical point of view in ML [611], even if some recent trends departs from this point of view [334].

Uncertainty management in KRR has a long history. It refers to the handling of incomplete information in non-monotonic reasoning as well as the handling of probabilities in Bayesian nets [494], and in probabilistic logic languages [540, 141]. However, the formalism of Bayesian nets is more appropriate for representing statistical data than incomplete data. The focus on uncertainty due to incomplete information is better captured by possibility theory (with weighted logic bases in possibilistic logic [207, 218]) and graphical representations (possibilistic nets [54, 59]). Belief functions also lend themselves to graphical representations (valuation networks [570], evidential networks [653]), and imprecise probability as well (credal nets [140]).

Uncertainty theories distinct from standard probability theory, such as possibility theory or evidence theory are now well-recognised in knowledge representation [185, 184]. They offer complementary views to uncertainty with respect to probability, or as generalisations of it, dedicated to epistemic uncertainty when information is imprecise or partly missing.

In KRR, at a more symbolic level, the inevitability of partial information has motivated the need for exception-tolerant reasoning. For instance, one may provisionally conclude that "Tweety flies" while only knowing that "Tweety is a bird", although the default rule "birds fly" has exceptions, and we may later conclude that "Tweety does not fly", when getting more (factual) information about Tweety. Thus non-monotonic reasoning [90] has been developed for handling situations with incomplete data, where only plausible tentative conclusions can be derived. Generic knowledge may be missing as well. For example, one may not have the appropriate pieces of knowledge for concluding about some set of facts. Then it may call for interpolation between rules [551].

When information is in excess in KRR, it may mean that it is just redundant, but it becomes more likely that some inconsistency appears. Redundancy is not always a burden, and may sometimes be an advantage by making more things explicit in different formats (e.g., when looking for solutions to a set of constraints).

Inconsistency is a natural phenomenon in particular when trying to use information coming from different sources. Reasoning from inconsistent information is not possible in classical logic (without trivialisation). It has been extensively studied in AI [63, 56, 104], in order to try and salvage non-trivial conclusions not involved in contradictions. Inconsistency usually appears at the factual level, for instance a logical base with no model. However, a set of rules may be said to be *incoherent* when there exists an input fact that, together with the rules, would create inconsistency [31].

ML can face several types of situations regarding the amount of information available. It must be said at once that induction, that goes from observations to regularities, is subject to the same kind of conservation law as in Physics. The information extracted is not created, it is just a reformulation, often with loss, of the incoming information.

If the input data is scarce, then prior knowledge, in one form or another, must complete it. The less data is available, the more prior knowledge is needed to focus the search of regularities by the learning system.

This is in essence what the statistical theory of learning says [611]. In recent years, lots of methods have been developed to confront the case where data is scarce and the search space for regularities is gigantic, specially when the number of descriptors is large, often in the thousands or more. The idea is to express special constraints in the so-called regularization term in the inductive criterion that the system use to search the hypothesis space. For instance, a constraint is often that the hypothesis should use a very limited set of descriptors [595].

When there is plenty of data, the problem is more one of dealing with potential inconsistencies. However, except in the symbolic machine learning methods, mostly studied in the 1980s, there is no systematic or principled ways of dealing with inconsistent data. Either the data is pre-processed in order to remove these inconsistencies, and this means having the appropriate prior knowledge to do so, or one relies on the hope that the learning method is robust enough to these inconsistencies and can somehow smooth them up. Too much data may also call for trying to identify a subset of representative data (a relevant sample), as sometimes done in case-based reasoning, when removing redundant cases. Regarding the lack of data there is a variety of approaches for the imputation of missing values ranging from the EM algorithm [175] to analogical proportion-based inference [85]. However these methods get rid of incompleteness and do not reason about uncertainty.

Finally, a situation that is increasingly encountered is that of multi-source data. Then, the characteristics of the multiple data sets can vary, both in the format, the certainty, the precision, and so on. Techniques like data fusion, data aggregation or data integration are called for, often resorting again to prior knowledge, using for instance ontologies to enrich the data.

2.4 Explainability and Causality

The need for explanations to enrich the prediction made by any AI model is not new and has been emphasized since the beginning of the Artificial Intelligence domain. With the massive success of machine learning, especially deep learning, in tackling complex data such as images, audio, video and texts, AI has achieved a high level of result quality in a wide variety of domains, including sensitive ones, e.g. related to health or justice. In such cases, even very accurate models may be insufficient, because the cost of an error can be huge, making it intolerable (see e.g. [98]). The need for explanation has thus become a major issue in the last decade or so. At the scientific level, this has led to the development of a research domain called eXplainable Artificial Intelligence, XAI for short, following the terminology first proposed in 2016 by DARPA [291].

Enriching the prediction with an explanation, viewed as a rationale for the prediction, can be seen as a tool that helps make an informed decision and determine whether the system should be trusted or is possibly making an error. Besides, applications to image data, for instance, have shown that deep learning sometimes have implicit and undesirable biases. In some situations these biases are patent, but, in other cases, errors are less obvious and require an understanding of how the model works. Explanations can then be seen as a tool aiming at correcting the model, explaining the decision to the user, and in some situations it can reveal that the model has undesirable biases and / or uses features protected by the law (such as, gender, political orientations, ...). However, it has been shown that it is possible to train a model in order to hide such biases from XAI methods, in particular when using local feature importance indices [194, 579]. Using certified, logically grounded explanations may solve this issue to some extent [328], as they should be harder to manipulate, yet this would remain to study.

The need for the explanation of results and the interpretability of models is amplified by the fact that the performances of the AI systems often come along with an increased complexity of the models, which makes them look like black boxes, opaque to the user. Moreover, in contrast with KRR models that may be also complex, the knowledge embedded in ML models is extracted from data, and as such offers no guarantee on the quality of its coverage. This has triggered the interest for explanation both in the KRR and ML communities.

Defining explainability

The interest in AI for explanations is not new. It already appears with the development of rule-based expert systems in the mid-1980's. At the time, there was a natural need for explanations that are synthetic, informative, and understandable for the user of an expert system [110]. This concern raises issues such as designing strategic explanations for a diagnosis, for example in order to try to lay bare the plans and methods used in reaching a goal [301], or using “deep” knowledge for improving explanations [368]. Another issue was the ability to provide negative explanations (not only positive ones) for answering questions of the form "Why did you not conclude X?" [534], even in the presence of uncertainty [239]. Let us also mention the problem of explaining the results of a multi-attribute preference model that is like a “black box”. It has been also studied in [387].

XAI has very fast given rise to a huge multiplicity of approaches and methods, but also a very high number of taxonomies to structure the domain. For some introductory surveys and discussions, see [73, 312, 314, 377, 313, 267, 287, 268, 454, 47, 467, 76].

"What is an explanation?", "What has to be explained?", and "how". These issues can be traced back to the absence of consensus on the definition of explanation that constitutes a multidisciplinary notion, related to psychology [594], philosophy [92, 542], cognitive sciences, education sciences or law, to name a few [621, 454]. It includes a human, subjective, component that is difficult to capture. Despite works on formal and axiomatic approaches [341, 12], there is no consensus on the definition of explanation that encompasses the variety of possible situations and the explainee perception (see below). The notion of a good explanation and of explanation quality is still a debated topic that gives rises to a multitude of rich discussions (see e.g. [199, 417, 315, 633] for some of them).

The breadth of the XAI domain can be illustrated by its diversity at several levels (see for instance the previously mentioned references [73, 287, 47, 467, 76] for more details): a *diversity of terms* can first be observed, with names such as explainability, interpretability, accountability or transparency. They cover related notions with somehow subtle nuances, but are sometimes used interchangeably. At a second level, a *diversity of tasks* can also be observed: the explanation may be requested for several types of machine learning tasks, such as classification, regression, clustering, outlier detection or recommendation but also, more broadly, for other artificial intelligence tasks such as planification, human-agent interactions or model conception.

At a third level, a *diversity of explainees* plays an important role: whether the explanation is generated for final users, domain experts or computer scientists, it needs to satisfy different requirements. On a related note, the aim of the explainees can differ, which can be summarised by the question they ask: it is often considered users want to know *why* a prediction is made, but they may also ask *how* they can get another prediction. The *how* question refers to the issue of so-called actionable explanations, i.e. that can lead to an action, and to the domain of algorithmic recourse [366, 365]. In all cases, it must be underlined that the aim is to explain an AI model, not the underlying ground truth: the explanation does not aim at providing information about reality, but on the AI model processes. As an example taken from [521], in the classification of images depicting dogs vs. wolves, finding that a decision was made because of the presence of snow in the background explains what are the conclusive evidence for the model, pointing to some issues in the training set it has been trained on.

At a fourth, more technical, level, explanation generation methods also differ in the *hypotheses* they rely on, for instance depending on what they consider as accessible: the model itself, or its type, training data or other, possibly unlabelled, data, to name a few. Such hypotheses include specifying whether expert or prior knowledge is available, for instance information about the descriptive features, their correlation or structural relations. Other possibilities include information about the users who receive the explanations, for instance their preferences, opening the way for personalised explanations.

Another issue concerns the statistical point of view according to which an interpretable model is a model that comes with mathematical guarantees [319, 23]. They are usually bounds for approximation errors (linked to the expression power of the hypothesis space) or the generalization error (linked to the robustness of the algorithm with respect to variations of the sample set). These can be also guarantees about the uncertainty around the parameters of the model (represented by confidence intervals for instance). Linear approaches are, in this scope, the most statistically interpretable ML algorithm. Robustness properties of statistical models are also desirable for interpretability.

Explainability and causality: a KRR point of view

The most developed topic of XAI in KRR, corresponds to the *diversity of explanation forms*; we shall see that that the same diversity also applies to the ML domain. The former favours formal approaches, in particular within a logical framework: knowledge-compilation based approaches represent AI models as Boolean circuits and consider explanations as prime implicants, i.e., minimal sufficient conjunctions to make a given statement true, e.g. attribute values that justify a given prediction [148, 150, 438]. A second type of approach exploits the formulation of explanations as an abductive reasoning issue [341]. Indeed, explanations are related to the question of identifying causes [300, 454, 455], i.e., to the notion of causality. Indeed most of the explanations we produce or we expect involve some causal relationships (e.g., John imposed on himself to go to the party *because* he thought that Mary would be there). In many domains where machines can provide aid for decision making, as in medicine, court decisions, credit approval and so on, decision makers and regulators more and more want to know what is the basis for the decision suggested by the machine, why it should be made, and what alternative decision could have been made, had the situation been slightly different. A question is: is it possible to extract causal relationships from data alone, without some prior knowledge that suggest those relationships? Judea Pearl [495, 496, 497] argues that this is not possible, but gives to the ML techniques the role of identifying possible correlations between variables in huge data sets that are impossible to sift through for human experts. A recent work [424] suggests that it would be possible to identify the direction of a causal relationship from observational data.

Causality is also closely related to the idea of counterfactual. Indeed an early definition of causality has been stated in terms of counterfactuals in a modal logic setting [620], and counterfactual is still a pervasive notion when discussing causality (even if causality cannot be reduced to the notion of a counterfactual) [214]. Counterfactual explanations for a decision process are translated into the notion of counterfactual examples [621, 286], that belong to the class of example-based explanations, together with prototypes, criticisms, influential instances (see e.g. [467]). They are built as follows. Given an example e, the counterfactual of e is the closest example to e (with respect to a metric) for which the decision changes (the counterfactual is not necessarily in the dataset). Consider for instance a model that determines if a credit is allowed or not with respect to the profiles of customers, and a given customer to whom the credit is not granted. The counterfactual in this case answers the question of what is the minimal change on his profile that would ensure that the credit is granted. If the model is based on propositional logic rules, the counterfactual will correspond to a minimal change of the considered example representation in Boolean logic [438]. In this case, the counterfactual is an understandable explanation for the prediction. In addition to the proximity of the counterfactual example to the instance of interest and the class prediction change, many other desiderata can be added into the cost function evaluating the quality of candidates, such as the sparsity of the change, its realism or its actionability, to name three examples: more than 60 counterfactual example generation methods have been proposed between 2015 and 2022 [286].

Machine learning explainability

The exploration of causality behind decisions in machine learning algorithms presents significant challenges. Machine learning primarily focuses on identifying correlations rather than causal relationships between input data and decisions. The case of identifying wolves versus dogs, as illustrated in [521], exemplifies this issue vividly. When employing a deep learning classifier, the most indicative feature for distinguishing between a dog and a wolf is the presence of snow. Although there is a noticeable correlation between snow and wolves, it is not indicative of a causal relationship in the ground truth. However, it is also indicative of a causal relationship in the prediction performed by the classifier. Initial research on explanations in machine learning concentrated on highlighting correlations between features and decisions. For example, Partial Dependence Plots [270] offer a graphical representation of feature correlations, while numerical approaches such as LIME [521] and SHAP [426] quantify the relative importance of each feature. These methods are considered as agnostic (and sometimes referred as "black-box," especially in the adversarial robustness area) meaning they analyze decisions inferred by the model without requiring direct access to the model internal machinery.

In this issue, [328] prove that the existing definitions of Shapley scores, often used in XAI methods, may *yield misleading information about the relative importance of features for predictions and the authors offer empirical evidence that such theoretical limitations of these scores are routinely observed in ML classifiers*

In the context of complex data and models, such as image classification in deep learning, defining explanations for specific decisions becomes increasingly relevant. Attribution methods aim to explain the prediction of a machine learning model by pointing out input variables that support the prediction – typically pixels or image regions for images – which lead to importance maps. Saliency [572] was the first proposed white-box (the architecture and the weights of the network are known) attribution method and consists of back-propagating the gradient from the output to the input. The resulting absolute gradient heatmap indicates which pixels affect the most the decision score. These methods were then followed by a plethora of other methods using gradients such as Integrated Gradient [589], SmoothGrad [580], Grad-cam [557] or Input Gradient [17]. All rely on the gradient calculation of the classification score. However, it is becoming increasingly clear that current methods raise many issues [7, 306, 577] such as confirmation bias: it is not because the explanations make sense to humans that they reflect the evidence on which the prediction is based.

Saliency maps can be considered as an attempt to provide counterfactual explanations. Essentially, the most effective method to identify the minimal change in an image that alters the decision of a classifier involves making slight adjustments in the direction of the model's gradient, which is precisely the way saliency maps are computed. However, within the realm of deep learning, counterfactual examples are related to the notion of adversarial examples [272], with the main difference that the latter typically involve an imperceptible alteration of the input. This suggests that gradient-based methods might reflect the structural vulnerabilities of networks rather than true causality. This perspective underscores the challenges in obtaining robust counterfactual explanations due to the inherent vulnerabilities of deep networks. Recent studies have concentrated on counterfactual explanation methods [622, 615], highlighting several desirable properties for these explanations [615]: Validity (ensuring the altered sample belongs to a different class), Actionability (a good counterfactual must not modify unchangeable features), Sparsity (minimize the number of features modified in the counterfactual), Data Manifold Closeness (counterfactual should be close to training data and respect the observed correlations between the features), and Causality (a counterfactual must maintain all known causal relationships between features). The latter three properties are typically not addressed by conventional adversarial attacks, prompting the development of more sophisticated methods [274, 529, 628]. Although validity may seem obvious, it is not trivial at all to ensure that with deep learning models. In [560], the authors demonstrate that with regularity constraints and a suitable loss function, saliency maps can encode counterfactual explanations through the lens of optimal transport[3].

XAI methods, from which the above paragraph mentions some examples, sometimes appear to be developed at a somehow theoretical level, without really taking into account the explainee, i.e. not considering the human dimension of the explanation receiver. As an illustration, it has been observed that many approaches lack experiments involving real users [372], with the need to measure separately their subjective satisfaction and their objective understanding [315]. In the case of attribution methods [416], authors propose to evaluate the alignment between these methods and human feature importance across 200,000 unique ImageNet images (called ClickMe dataset). The alignment between DNN Saliency and human explanations is quantified using the mean Spearman correlation, normalized by the average inter-rater alignment of humans. Good alignment scores with respect to human annotations does not guarantee that the explanation is satisfactory, but it has been shown that robust models tend to obtain better scores [560, 241]. In the same way, recent developments focus on Human-Centered eXplainable AI (see e.g. [410]) and on the question of how to display the outputs of the XAI methods, through appropriate interfaces, leading to the domain of eXplanation User Interfaces, XUI (see e.g. [119]).

The emergence of large generative models, such as large language models [93], complicates the task of explaining machine learning methods. Unlike models focusing on decision-making, generative models are

[3] This is a problem of optimal transportation and allocation of resources, originally formulated by Gaspard Monge, in a context of cut and fill. In its modern reformulation, it amounts to finding a transport map between two spaces equipped with probability measures. Then the cost of the optimal transport map is measured by the Wasserstein p-distance (for $p = 1$) between the two probability measures (the cost of a move between two points is simply the distance between them, in case the two spaces are identical). In the context of XAI, it involves constructing a counterfactual explanation for an instance of one class by identifying its counterpart according to the optimal transport map between classes [560].

centered on generation, necessitating innovative approaches to frame the question of explanation [77].

As announced at the end of Section 1, we have organized in seven sections the main areas where synergies between KRR and ML take place. We start with the learning of rule-based models which are at the core of induction tasks.

3 Rule-based models and neural-symbolic AI

Finding logical rules that cover a set of examples and exclude a set of counter-examples is a matter of induction.

Version space learning [461, 462] has offered a first framework for describing how a hypothesis space is organized. A hypothesis is understood as a concept, and takes the form of a tuple of values in attribute domains, including question marks when there is no restriction on some attribute values. Version space learning takes into account examples and counterexamples of the learned concept, maintaining upper and lower approximations of its representations (including examples and excluding counterexamples).

Inductive Logic Programming [567] attempts to learn rules from examples and background knowledge in the representation setting of logic programming. This is the topic of the next section with its extensions to probabilistic rules. Then this section ends with more recent neural-symbolic approaches, which aims at interfacing logic and probabilities with neural networks.

Note that the learning of non conventional rules such as default rules, rules with threshold, etc. is addressed in Section 6.

3.1 Inductive Logic Programming and statistical relational learning

Inductive Logic Programming (ILP) (see [473, 165, 224] for general presentations) is a subfield of ML that aims at learning models expressed in (subsets of) First Order Logic. It is an illustration of Symbolic Learning, where the hypothesis space is discrete and structured by a generality relation. The aim is then to find a hypothesis that covers the positive examples (it is then said to be complete) and rejects the negative ones (it is said to be consistent). The structure of the hypothesis space allows to generalize an incomplete hypothesis, so as to cover more positive examples, or to specialize an inconsistent hypothesis in order to exclude negative covered examples. The main reasoning mechanism is induction in the sense of generalization (subsumption).

In ILP, examples and models are represented by clauses. Relying on First Order Logic allows to model complex problems, involving structured objects (for instance to determine whether a molecule is active or not, a system must take into account the fact that it is composed of atoms with their own properties and shared relations), or involving objects in relation with each other (a social network or temporal data). Reasoning is a key part of ILP. First, the search for a model is usually performed by exploring a search space structured by a generality relation. A key point is then the definition of a generality relation between clauses. The more natural definition of subsumption should be expressed in terms of logical consequences, which allows comparing the models of both formula, but since the problem is in general not decidable, the notion of θ-subsumption, as introduced in [502] is usually preferred: a clause C_1 is more general that a clause C_2 if there exists a substitution θ such that $C_1.\theta \subseteq C_2$. In this definition a clause, i.e., a disjunction of literals, is represented by its set of literals. For instance, the rule $par(X,Y), par(Y,Z) \rightarrow grand\_par(X,Z)$ θ-subsumes $par(john, ann), par(ann, peter), par(john, luc) \rightarrow grand\_par(john, peter)$. Indeed, the first one leads to the clause $\neg par(X,Y) \vee \neg par(Y,Z) \vee grand\_par(X,Z)$ and the second one to the clause $\neg par(john, ann) \vee \neg par(ann, peter) \vee \neg par(john, luc) \vee grand\_par(john, peter)$. Second, expert knowledge can be expressed using facts (ground atoms) or by rules, or yet reasoning mechanisms to be applied. This can be illustrated by the well-known systems FOIL [513] and Progol [472].

ILP, and more generally Symbolic Learning, has thus some interesting properties. First, the model is expressed in logic and therefore is claimed to be easily understandable by a user (See for instance [474] for an interesting study of the comprehensibility or not of programs learned with ILP). Second, expert knowledge can be easily expressed by means of clauses and integrated into the learning algorithm. Although initially developed for the induction of logic programs, it has now shown its interest for learning with structured data.

However, ILP suffers from two drawbacks: the complexity of its algorithms and its inability to deal with uncertain data. Several mechanisms have been introduced to reduce the complexity, as for instance the introduction of syntactic biases, restricting the class of clauses that can be learned. Another interesting idea is propositionnalization, introduced in [393] and then developed for instance in the system RSD [668]. It is a process that transforms a relational problem into a classical attribute-value problem by the introduction of new features capturing relations between objects. Once the transformation performed, any supervised learner can be applied to the problem. The main difficulty is then to define these new features.

This last problem has led to the emergence of Statistical Relational Learning [264, 167, 170] that aims at coupling ILP with probabilistic models. Many systems have been developed, extending naive Bayesian classifier [388], Bayesian Networks [245] or Markov Logic Networks [523] or developing new probabilistic framework as in Problog [171], or in "Probabilistic Soft Logic (PSL)" [40]. This latter work belongs to an approach that combines probabilities with fuzzy logic connectives [234, 237]. In all these works, inference and learning are tightly connected since learning parameters requires to maximize the likelihood for generative learning (estimation of the probabilities to generate the data, given a set of parameters), or the conditional likelihood in case of discriminative learning (estimation of the probabilities of the labels given the data). Optimizing the parameters thus requires at each step to estimate the corresponding probabilities. This has led to intensive research on the complexity of inference. A comparative study of weight learning methods for first-order logical rules can be found in [585].

3.2 Neural-Symbolic Reasoning

Neural-Symbolic AI is the banner of an important research trend nowadays that aims at putting together logic, probability and neural nets [435]. Several works have proposed to combine learning and reasoning by studying schemes to translate logical representations of knowledge into neural networks. A long-term goal of a series of works on neural-symbolic integration, surveyed for instance by [64], is "to provide a coherent, unifying view for logic and connectionism ... *[in order to]* ... produce better computational tools for integrated ML and reasoning." Existing works can be distinguished based on whether they only involve reasoning (e.g. using neural networks as a faster alternative to symbolic reasoners) or also learning, and based on whether uncertainty is explicitly taken into account.

Neural-Symbolic AI for Reasoning The problem of interfacing logical representations with machine learning devices was already addressed in the 1990s. Early works proposed translation algorithms from a symbolic to a connectionist representation, enabling the use of computation methods associated with neural networks to perform symbolic reasoning. Works in this vein include [499, 500, 79, 317]. Bornscheuer *et al.* [79] show for example how an instance of the Boolean satisfiability problem can be translated into a feed-forward network that parallelizes GSAT, a local-search algorithm for Boolean satisfiability. They also show that a normal logic program P can be turned into a neural network that can approximate the semantics of the program arbitrarily well. This kind of translation has been also proposed for non-classical logics. In [160, 157], methods are proposed to translate formulas with modalities into neural networks, enabling the representation of time and knowledge, and the authors in [158, 159] show that there exists a neural network ensemble that captures fixed-point semantics of intuitionistic theories. The idea of using neural networks for reasoning has remained an active area of research. See, for instance, [309] for a recent survey. Rather than relying on manual translations, however, the emphasis is now on training neural networks to approximate the results of a symbolic reasoner. The main motivation is to enable fast (albeit approximate) inference results and, potentially, increased robustness in the presence of noisy inputs.

Neural-Symbolic AI for Learning and Reasoning Short after the beginning of the first wave of works mainly focusing on the translation of logic programs into neural networks, many works in neuro-symbolic AI became more ambitious, by incorporating a genuine learning step. Towell and Shavlik [602] offer one of the first systems, named KBANN (for Knowledge-Based Artificial Neural Networks), where both reasoning and learning are jointly at work. They wrote, "Briefly, the idea is to insert a set of hand-constructed, symbolic rules (i.e., a hand-built classifier) into a neural network. The network is then refined using standard neural learning algorithms and a set of classified training examples. The refined network can then function as a highly-accurate classifier." [601] presents a "final step for KBANN, [that is] the extraction of refined, comprehensible rules from the trained neural network". In a similar spirit, the paper [162] represents propositional logic programs with recurrent neural networks (RNNs) which can be used to compute the semantics of the program. They show that this program can also be used as background knowledge to learn from examples, using back-propagation. Essentially, the RNN defined to represent a logic program P has all atoms of P in the input layer; one neuron, a kind of "and" gate, for each rule in a single hidden layer; and one neuron for every atom in the output layer, working like "or" gates. Re-entrant connections from an atom in the output layer to its counterpart in the input layer enable the chaining of rules. An extraction step is also described in [154]. Franca *et al.* [250] extend these results to first-order programs, using a propositionalization method called Bottom Clause Propositionalization. Pinkas and Cohen [501] performed experiments with so-called higher-order sigma-pi units (which compute a sum of products of their inputs) instead of hidden layers, for planning problems on simple block-world problems. The number of units is fixed at design time, and is a function of the maximum number of blocks and the maximum number of time steps. For example, for every pair (b_1, b_2) of possible blocks and every time step t, there is a unit representing the proposition above(b_1, b_2, t). Their results indicate that a learning phase enables the network to approximately learn the constraints with a reasonable number of iterations.

The aforementioned works use special-purpose neural network architectures, whose structure remains closely aligned with symbolic rule bases. Another direction of research has emerged, in which more general neural network architectures are used, and where the aim is merely to train the network to learn to reason about a particular domain, which is particularly popular in the context of knowledge graph completion. For instance, in [654, 316], they use recursive tensor networks to predict classes and / or binary relations from a given knowledge base. A popular strategy is to encode prior symbolic knowledge, when training the neural network, by interpreting the logical connectives in terms of fuzzy logic connectives [193, 581, 558]. The main motivation for using fuzzy logic connectives, in this context, is that this makes it possible to translate symbolic knowledge into a continuous regularisation term in the loss function. Donadello *et al.* [197] describe how this approach can be used to learn semantic image interpretation using background knowledge in the form of simple first-order formulas. Rather than directly regularizing the loss function in this way, [323] proposes an iterative method to ensure that the proportion of ground instances of the given rules that are predicted to be true by the neural network is in accordance with the confidence we have in these rules. To this end, after each iteration, they solve an optimisation problem to find the set of predictions that is closest to the predictions of the current neural network while being in accordance with the rules. The neural network is subsequently trained to mimic these regularized predictions. Yet another approach is proposed in [648], which proposes a loss function that encourages the output of a neural network to satisfy a predefined set of symbolic constraints, taking advantage of efficient weighted model counting techniques. *In this issue, [269] proposes a novel neuro-symbolic framework integrating propositional logic requirements into the output layer of neural networks, ensuring compliance with the requirements and enhancing performance. Extensive experimental evaluation shows that CCN+ outperforms both its neural counterparts and the state-of-the-art models in multi-label classification tasks.* All of these methods are essentially aimed at incorporating background knowledge.

Another line of work is aimed at improving the reasoning capabilities of neural networks. For instance, the Neural Theorem Prover (NTP) [456] is essentially a logic programming framework, where the usual unification between terms is replaced by a soft unification mechanism, which depends on the similarity between embedding representations. This framework makes it possible to start from rule templates, which only specify the structure of the rules to be learned. The model then learns representations of the predicates appearing in this templates, thus instantiating the templates with specific (soft) rules. Lifted Relational Neural Networks [582] follow the same principle, but rather than relying on embeddings, they rely on latent predicates with learned membership degrees.

Neural Probabilistic Reasoning Statistical relational learning is concerned with combining logic with probabilities, whereas neuro-symbolic AI is concerned with combining logic with neural networks. A recent research trend is focused on the joint combination of neural networks, logic and probabilities. A seminal work in this area is DeepProbLog [434], which relies on probabilistic logic programs where the probabilities of facts are predicted by a neural network. The key innovation relates to how the neural network can be learned, using backpropagation, given a loss function that is defined in terms of the consequences of the probabilistic logic program. *In this issue, [190] provides an overview and synthesis of neuro-symbolic methods, thereby contributing a unified algebraic perspective on the different flavors of probabilistic logic programming (PLP), showing that many if not most of the extensions of PLP can be cast within a common algebraic logic programming framework, in which facts are labeled with elements of a semiring and disjunction and conjunction are replaced by addition and multiplication. In this unified perspective, the authors focus on the ProbLog language and its extensions.* Other relevant work in this area includes NeuPSL (for Neural Probabilistic Soft Logic)[510], NeurASP [679], TensorLog [123] and neural Markov logic [444]. *In this issue [356] proposes a new class of Neural Markov Logic Networks (NMLN), called Quantified NMLN, that extends the expressivity of NMLNs and demonstrate how to leverage the neural nature of NMLNs to employ learnable aggregation functions as quantifiers, and demonstrate the efficiency of Quantified NMLNs in molecule generation experiments.* Broadly speaking, the main aim of such approaches is to have a close cooperation between perception and recognition (as provided by a neural net), and reasoning capabilities. This can be exemplified by the exploitation of constraints linking symbols that are recognised, as for instance in the disambiguation of additions of handwritten figures [434], or in the validation of handwritten sudoku puzzles [627].

Comparing StarAI and Neural-Symbolic Approaches The detailed surveys [166, 443] compare several works in STAtistical Relational Artificial Intelligence and Neural-Symbolic Learning and Reasoning along several dimensions. In particular, they emphasize the following dimensions: i) one can distinguish between symbolic representations and sub-symbolic representations, such as pixels in an image or vector embeddings of predicates viewed as words ; ii) while statistical relational learning relies on both logic and probabilities, neural-symbolic approaches use fuzzy logic connectives in order to have a differentiable representation liable to be handled in machine learning. Thus neural net outputs are interpreted as probability distributions in statistical relational learning, and often as a fuzzy set in neuro-symbolic approaches.

4 Using background knowledge in learning

It is worth taking a historical perspective on machine learning for a moment. Apart from the problem of learning from "pre-classified learning examples" (as presented in [568] published in 1990), which included version space learning, the multi-layer perceptrons and decision trees, early learning systems were closely associated with problem solving, and problem solving itself with reasoning or planning systems. The aim was to improve problem-solving efficiency by drawing on past experience, or on advice or solutions provided by experts. One approach, called "explanation-based learning", attracted a great deal of interest in the 1980s. The idea is to consider a given solution to a particular problem and derive from it a general rule capable of solving as broad a class of problems as possible that includes the particular problem used as input. To do this, the learner needs to find an explanation of why the particular solution works, and generalize this explanation as far as possible so that it applies to similar problems that may arise in the future. Generalization relies on the availability of a theory of the domain expressed using some form of logic. Most methods have generalized the constants of the explanation into constrained variables, sometimes also generalizing the structure of the graph representing the explanation. Thanks to this generalized rule, the system can quickly solve related problems. This method has been applied to learning general rules for recognizing concepts ([464, 172]), macro-operators in the case of planning systems ([457]) and control rules or heuristics to guide the search for a solution in the vast tree of possible solutions ([458]).

In this perspective, the performance criterion was not the error rate, as would later be the case, but the gain in problem-solving speed. One of the limitations of the approach was the need for perfect or near-perfect (i.e. complete and consistent) domain theory, which seriously hampered its application to poorly known domains, whereas the emerging second connectionism [537] and, almost simultaneously, techniques such as support vector machines (SVMs) and boosting, were capable of handling learning data marred by noise and even missing values. However, as mentioned in Section 2.1, training data must be supplemented with prior knowledge. In the case of non-reasoning systems such as neural networks, SVMs or boosting, learning becomes an optimization process that seeks an optimum of a certain criterion combining the fit to the training data and satisfaction of constraints that express prior knowledge.

These constraints can be expressed as:

- regularization terms directly incorporated in the criterion to be optimized
- ways of modifying the training data themselves by modifying their representation

Although prior knowledge is thus limited to a somewhat impoverished form, it is easily integrated into the essentially digital optimization perspective that learning has become, hence its generality and power. However, the success of learning in the big data area has also led to a new demand. This is to be able to provide explanations of the results and/or the learning process itself. This new development is attracting a great deal of attention today, and could lead to renewed interest in hybridization with symbolic reasoning systems. Time will tell. In parallel with this new demand for explainable AI, it is very likely that error rate as the sole measure of performance will give way to richer, more encompassing criteria, a trend that could itself bring new learning methods to the fore. To illustrate this growing interest, the article [62] considers the question "What could serve as a training objective for mathematical discovery?" and advocates a look at the usefulness of existing theorems and the one conjectured as one criterion, among others, to guide an "AI mathematician" in its exploration in the space of all possible mathematical statements.

This section describes ways of incorporating prior knowledge in learning.

4.1 Regularization and search biases

Attempting to induce general laws from limited training data without limiting the space of hypotheses (the laws) is bound to produce, except very improbably by pure chance, hypotheses that have no values. In particular, the phenomenon of overfitting occurs when the hypothesis produced fits too well the training data and their specific but irrelevant characteristics, which is likely if the hypothesis space has too great a capacity. This is well explained by the statistical theory of learning [611]. Thus, in order to complement the training data, it is necessary to use some prior knowledge that restricts the space of hypotheses to be explored.

Generally, two forms of prior knowledge, aka. biases, are distinguished: *representation biases* that limit the expressiveness of the language used to express the possible hypotheses on the world, and *search biases* that control how the hypothesis space is explored by the learning algorithm.

Representation biases can take various forms. They can directly affect the language in which the possible hypotheses can be expressed. For instance, "hypotheses can involve a maximum of two disjuncts". Or, even if the useful features are present in the training database (e.g. 'distance' and 'time'), it may help the learning system if a feature that combine them (e.g. 'speed') is added. In this case, the expert can prove very useful. Recently, we have seen the emergence of learning systems with millions or billions of parameters, a number that easily exceeds the number of learning examples available, thus creating a very high risk of over-fitting. In this context, one way of biasing the system is to learn what is called an "embedding", i.e. to learn an intermediate representation from the raw input, where it is easier to discern patterns of interest and/or to extract prediction rules. In this perspective, the role of the expert is no longer to directly shape the input space, but to find ways of directing the self-supervised learning process that produces these embeddings. For instance, in contrastive learning, the use of siamese networks will help encode prior knowledge about similar training examples and examples that should be distant.

A more sophisticated approach, used in deep neural networks, is to train the system while respecting known laws about the phenomenon at hand, or other kinds of expert knowledge. A generic way of imposing representation biases is thus to use a regularized optimisation criterion that balances a measure of fit of the model to the data, and a measure of fit of the model to the bias. For instance, the following quality measure over linear hypotheses $h(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}$ for regression favours hypotheses that involve fewer parameters:

$$R(h) = \underbrace{\frac{1}{2}\sum_{i=1}^{m}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{i,j}\right)^2}_{\text{fit to the data}} + \underbrace{\lambda \sum_{j=1}^{p} ||\beta_j||_0}_{\text{Favors models with few non zero parameters}}$$

where the L_0 norm $||.||_0$ counts the nonzero parameters β_j.

Thus, for instance, Physics-Informed Neural Networks (PINNs) attempt both to optimize the fit to the training data (e.g. as measured by the Mean-Square Error) and to satisfy initial and boundary conditions imposed by equations (e.g. differential equations) that express knowledge of the physical laws governing the phenomenon [518, 509]. PINNs seek to minimize a combined fit to the data and the satisfaction of physics based conditions, albeit with the help of specific architectures for the neural networks, e.g. with two branches that are constrained to share parameters, one dedicated to the fit of the training data and the other to the representation of the physics involved. We have seen in Section 3.2 how the use of fuzzy connectives enables the expression of some background knowledge in a regularization term.

The *search bias* dictates how the learning algorithm explores the space of hypotheses. For instance, in the case of neural networks, the search starts with a randomly initialized neural network and then proceeds by a gradient descent optimization scheme. In some other learning methods, such as learning with version space, the search uses generalization relations between hypotheses in order to converge towards good hypotheses. In this latter case, it is easier to incorporate prior knowledge from the experts. Indeed, the exploration of the hypothesis space is akin to a reasoning process, very much like theorem proving.

It should be noted that in transfer learning (see Section 5.3), the source task acts as a bias for the target task by providing a starting model that must be adapted to the target domain.

While a bias is necessary in inductive learning, it is important to realize that it can help as well as hinder the discovery of a good model. However prior knowledge is expressed, through a representation bias and/or a search bias, its appropriateness is the responsibility of the domain expert and the data scientist.

4.2 Aligning Symbolic and Vector Space Representations

Symbolic and vector representations clearly have some complementary strengths for encoding knowledge. For instance, vector representations are typically easier to learn due to their continuous nature. Moreover, vector representations can model certain aspects of knowledge in a more fine-grained way, e.g. by capturing degrees of similarity or the intensity with which some property is satisfied. Symbolic representations, on the other hand, are typically easier to interpret, which makes models that rely on symbolic representations also easier to explain. Similarly, in many domains we have access to knowledge bases that are encoded in symbolic form (e.g. using logic or natural language). In this section, we discuss methods that align symbolic and vector representations in some way, in attempt to combine the best of both worlds.

Conceptual Spaces The problem of aligning symbolic and vector space representations lies at the heart of the theory of *conceptual spaces* [262], which was proposed by Gärdenfors as an intermediate representation level between vector space representations and symbolic representations. Conceptual spaces are essentially vector space models, as each object from the domain of discourse is represented as a vector, but they differ in two crucial ways. First, the dimensions of a conceptual space usually correspond to interpretable salient features. Second, (natural) properties and concepts are explicitly modelled as (convex) regions. Given a conceptual space representation, we can thus, e.g., enumerate which properties are satisfied by a given object, determine whether two concepts are disjoint or not, or rank objects according to a given (salient) ordinal feature. Conceptual spaces were proposed as a framework for studying cognitive and linguistic phenomena, such as concept combination, metaphor and vagueness. As such, the problem of learning conceptual spaces from data has not yet received much attention. Within a broader setting, however, several authors have studied approaches for learning vector space representations that share important characteristics with conceptual spaces. The main focus in this context has been on learning vector space models with interpretable dimensions. For example, it has been proposed that non-negative matrix factorization leads to representations with dimensions that are easier to interpret than those obtained with other matrix factorization methods [398], especially when combined with sparseness constraints [320]. More recently, a large number of neural network models have been proposed with the aim of learning vectors with interpretable dimensions, under the umbrella term of *disentangled representation learning* [113, 307]. Another possibility, advocated in [191], is to learn (non-orthogonal) directions that model interpretable salient features within a vector space whose dimensions themselves may not be interpretable. Beyond interpretable dimensions, some authors have also looked at modelling properties and concepts as regions in a vector space. For example, [226] proposed to learn region representations of word meaning. More recent approaches along these lines include [618], where words are modelled as Gaussian distributions, and [351], where word regions were learned using an ordinal regression model with a quadratic kernel. Some authors have also looked at inducing region based representations of concepts from the vector representations of known instances of these concepts [87, 292]. Finally, within a broader setting, some approaches have been developed that link vectors to natural language descriptions, for instance linking word vectors to dictionary definitions [308] or images to captions [367].

An important question when dealing with embeddings is how much they can be used to perform logical reasoning. In this issue [550] studies to which extent and under which conditions pooling operators, that are commonly used for aggregating vector embeddings in deep neural network architecture, are compatible with the idea that embeddings encode epistemic states and fit with the satisfaction of propositional formulas. The paper shows that max-pooling is particularly suitable for such tasks, in particular if one wants to encode non-monotonic reasoning.

Concept Bottleneck Models A popular strategy for building interpretable classifiers is based on aligning representations learned by deep neural network models with interpretable concepts. For instance, Concept Bottleneck Models [378] approach image classification in two steps. First a traditional image classification model is used to predict which primitive concepts can be observed in an image. Second, a final prediction is made based on the primitive concepts that were identified in the first step. For instance, when classifying images of birds, the concepts involved may correspond to features such as the colour of the wings. The model would then use such features to predict the species of the bird that is depicted. Concept Bottleneck Models offer a degree of interpretability. For instance, we can easily inspect the primitive concepts that were identified to analyse why a given image was misclassified. Moreover, we may expect that making predictions in terms of primitive concepts can also serve as a form of semantic regularization. However, concept bottleneck models require appropriate supervision data, to allow them to learn the primitive concepts as well as the target classes, which is not always available in practice. Moreover, they are only useful in cases where an exhaustive set of primitive concepts can be identified a priori. This is often difficult to achieve in practice, leading to models with sub-optimal performance [227]. Furthermore, it has been observed that the representations learned by concept bottleneck models may capture information beyond the considered primitive concepts, making the resulting interpretations potentially misleading [428]. As an alternative to concept bottleneck models, [661] suggest to learn concept-based explanations as a post-hoc step. Another alternative is to design classifiers that can be explained by comparing instances to learned prototypes [404]. Such approaches avoid the need for pre-defined concepts, but the learned prototypes may not always be easily interpretable. Some methods inspired by conceptual spaces have been proposed as well. For instance, [191] extract an interpretable (qualitative) representation from a given vector space, by essentially identifying directions within the space that correspond to interpretable properties. To train an interpretable classifier, they simply treat these properties as ordinal features.

4.3 Using Knowledge Graphs for Learning

Knowledge graphs (KGs) are a popular formalism for expressing relational knowledge using triples of the form (entity, relation, entity). In application fields such as natural language processing, they are among the most widely used knowledge representation frameworks. For instance, several authors have explored strategies for incorporating KGs when training language models [675, 421, 629]. KGs are also commonly used to provide background knowledge in NLP tasks such as question answering. For instance, commonsense KGs such as ConceptNet[4] and ATOMIC [546] are commonly used to improve the commonsense reasoning abilities of language models [658, 674]. In computer vision, commonsense knowledge graphs have similarly been used to interpret visual scenes [285, 645]. While recent Large Language Models, such as ChatGPT and GPT-4, capture a wealth of world knowledge, they are still limited when it comes to lesser-known entities [432]. The extent to which they can be kept up-to-date is also inherently limited. KGs thus still have an important role to play in providing background knowledge to language models. KGs have also found important applications in the field of recommendation [670, 625, 630]. KGs can improve the quality of recommendations, by giving the system side information about the relatedness of different items, but

[4] https://conceptnet.io

they are also useful for making the recommendation process more explainable [631]. The usefulness of KGs for explainable machine learning more generally has also been highlighted [596].

Knowledge Graph Completion Knowledge graphs are almost inevitably incomplete, given the sheer amount of knowledge about the world that we would like to have access to and given the fact that much of this knowledge needs to be constantly updated. This has given rise to a wide range of methods for automatic knowledge graph completion, which is clearly a research area that has the integration of background knowledge and machine learning at its heart. On the one hand, several authors have proposed approaches for automatically extracting missing knowledge graph triples from text [524]. On the other hand, a large number of approaches have been studied that aim to predict plausible triples based on statistical regularities in the given knowledge graph. Most of these approaches rely on vector space embeddings of the knowledge graph [78, 603]. The main underlying idea is to learn a vector $\mathbf{e}$ of typically a few hundred dimensions for each entity e, and a scoring function s_R for each relation R, such that the triple (e, R, f) holds if and only if $s_R(\mathbf{e}, \mathbf{f}) \in \mathbb{R}$ is sufficiently high. Provided that the number of dimensions is sufficiently high, any knowledge graph can in principle be modelled exactly in this way [371]. To a more limited extent, such vector representations can even capture ontological rules [293]. In practice, however, our aim is usually not to learn an exact representation of the knowledge graph, but to learn a vector representation which is predictive of triples that are plausible, despite not being among those in the given knowledge graph. Some authors have also proposed methods for incorporating textual information into knowledge graph embedding approaches. Such methods aim to learn vector space representations of the knowledge graph that depend on both the given knowledge graph triples and textual descriptions of the entities [676, 350, 643, 642], or their relationships [525, 600]. Finally, several authors have started to explore how KGs can be extracted directly form Large Language Models [635, 122].

In this issue, [538] proposes techniques to generate explanations for predictions based on the embeddings of knowledge graphs: the idea is to build explanations out of paths in an input knowledge graph, searched through contextual and heuristic cues. On their side, [102] presents a model for fuzzy temporal reasoning to overcome some inconsistencies detected in pre-trained language models in a specific application domain of a conversational agent carefully designed for providing users with explanations. More precisely, starting from a knowledge graph that offers an intuitive representation of the entities and relations in the application domain, the authors describe how to map the temporal information onto a fuzzy temporal constraint network. An experiment with GPT-3 Large Language Model is reported.

5 Reasoning for learning by using constraints, semantic features, analogies

The idea of combining reasoning and learning covers many issues. In Section 3, we already mentioned version space learning, and we surveyed ILP, statistical relational learning, and neural symbolic AI. In that section, we saw some interplay between logic programming techniques, probabilistic reasoning, or fuzzy logic and neural networks. In Section 4, background knowledge was used either in the bias in order to improve the learned model, or to learn a new representation space.

This section covers machine learning settings in which various forms of reasoning play a central role. This involves reasoning about the predictions of machine learning models, the use of reasoning to satisfy declarative constraints, and the use of special forms of inference to make predictions.

More specifically, this section successively deals with:

- Declarative biases for pattern mining and clustering.
- Low-shot learning where knowledge is used in order to be able to learn from very few training examples.
- Transductive learning methods such as case-based reasoning and analogical reasoning, and their relation with transfer learning.

The combination of logical representations with learning is also considered in Section 6 (learning for knowledge acquisition), in Section 7 (learning for reasoning), where the focus is on learning representations that can be used for reasoning. Logical representations, especially in the form of rules, can also be learned in the setting of explainable AI, as covered in Section 8 (model accountability).

5.1 Declarative Frameworks for Data Mining and Clustering

Unsupervised ML tasks such as clustering can be guided by expert knowledge expressed by means of constraints that can the handled from declarative frameworks developed in the KRR side. Machine Learning and Data Mining can also be studied from the viewpoint of problem solving. From this point of view, two families of methods can be distinguished: enumeration and optimisation, the latter being either discrete or continuous.

Pattern mining is the best known example of enumeration problems, with the search for patterns satisfying some properties, as for instance being frequent, closed, emergent ... Besides, supervised classification is seen as the search for a model minimizing a given loss function, coupled to a regularization term to avoid over-fitting, whereas unsupervised learning is modeled as the search of a set of clusters (a partition in many cases) optimizing a quality criterion (the sum of squared errors for instance for k-means). To cope with complexity, optimisation problems usually rely on heuristic search, with the risk of finding only a local optimum. All these approaches suffer from drawbacks. For instance in pattern mining the expert is often overwhelmed by all the patterns satisfying the given criteria. In optimisation problems, a local optimum can be far from the expert expectations.

To prevent these drawbacks, the notion of Declarative Data Mining has emerged, allowing the experts to express knowledge in terms of constraints on the desired models. It can be seen as a generalization of semi-supervised classification, where some data are already labelled with classes. Classical algorithms must then be adapted to take into account constraints, which has led to numerous extensions. Most of them are dedicated to only one type of constraints, since the loss function has to be adapted to integrate their violation and the optimization method (usually a gradient descent) has to be adapted to the new optimization criterion. It has been shown in [168] that declarative frameworks, namely Constraint Programming in that paper, allow to model and handle different kinds of constraints in a generic framework, with no needs to rewrite the solving algorithm. This has been applied to pattern mining and then extended to k-pattern set mining with different applications, such as conceptual clustering or tiling [169, 375].

This pioneering work has opened the way to a new branch of research, mainly in Pattern Mining and in Constrained Clustering. In this last domain, the constraints were mainly pairwise, e.g., a Mustlink (resp. Cannotlink) constraint expresses that two points must (resp. must not) be in the same cluster. Other constraints have been considered such as cardinality constraints on the size of the clusters, minimum split constraints between clusters. Different declarative frameworks have been used, as for instance SAT [153, 348], Constraint Programming [144], Integer Linear Programming [470, 39, 382, 487]. An important point is that such frameworks allow to easily embed symbolic and numerical information, for instance by considering a continuous optimisation criterion linked with symbolic constraints, or by considering two optimisation criteria and building a Pareto front [144].

Thus, declarative frameworks not only allow to easily integrate constraints in Machine Learning problems, but they enable the integration of more complex domain knowledge that goes beyond classical Machine Learning constraints, thus integrating truly meaningful information [145]. Moreover, new use case for clustering can be considered as for instance given a clustering provided by an algorithm, find a new clustering satisfying new expert knowledge, minimally modifying the previous one [382]. The price to pay is computational complexity, and the inability to address large datasets. A new research direction aims at studying how constraints could be integrated to a deep learner [671, 480]. Besides, the Constraint Programming community has also benefited from this new research direction by the development of global constraints tailored to optimize the modeling of Data Mining tasks, as for instance [373].

A detailed survey of declarative frameworks for clustering can found in this issue [146].

5.2 Low-shot learning

In this subsection, we point out that low-shot learning, a specific form of learning with few or no examples, encounters KRR in two respects: by requiring some form of knowledge describing classes, and by being close to the concerns of case-based reasoning. A recent survey can be found in [322].

One of the main cognitive tasks is to identify the category to which an object such as a pattern in an image, a sound, a text, etc. belongs. In Machine Learning, this problem involves learning a function that associates an input $\mathbf{x}$ (e.g., a pattern) with a class or category y (e.g., "a car"). In the standard supervised learning scenario, this is achieved by presenting a set of m (m being large, or even very large) training examples $\{(\mathbf{x}_i, y_i)\}_{1 \leq i \leq m}$ to a learning algorithm and learn a recognition rule for each class.

But we humans often do not need such large training datasets to learn how to classify objects. For example, children learn to recognize cats from very few examples, and if given a definition of "zebra" (i.e. a wild African horse with black and white stripes and an erect mane), they are even able to recognize them from pictures without ever having seen one before. These two situations, quite common in natural cognition, have been called "learning from a few images" and "learning from zero images" respectively. But how are such feats even possible?

Standard supervised learning does not involve any reasoning. The classes y_i are arbitrary tokens for the learning system. The class "cat" could as well be referred as "QP115", that would not change in any way the learning process. The class tags are unrelated and the system attempts to capture correlations between the description of the inputs $\mathbf{x}$ and the class y to which they are paired in the training dataset without taking into account any semantics associated with the objects and classes nor any relationships between the classes.

If you have never been exposed to an instance of a class, the only way to be able to recognize its occurrence is to relate its description to other classes or to examples. For instance, if you know how to recognize a "horse" and a "striped" object, then you may be able to recognize a "zebra" from a new example because it triggers both the "horse" and "stripped" categories. The overall idea then is to describe the classes as compositions of *semantic features*, and learn to recognize these features when presented with an example. This is somewhat reminiscent of the use of background knowledge as described in subsection 4.2 where some intermediate representation level between vector space representations and symbolic representations is used. In both cases, disentangled clusters corresponding to concepts are looked for in the intermediate representation space. But in the conceptual space approach, an interpretation in terms of symbolic knowledge is aimed at in order to study cognitive phenomena, whereas in zero-shot learning, the activation of several "concepts" is just a means to recognize a new type of pattern, in a first approach irrespective of any symbolic interpretation beside this recognition.

Learning to associate examples to classes then becomes a two step process, the first being to learn the semantic features $F \in \mathscr{F}$ allowing the description of the classes, and then to learn to associate examples $\mathbf{x}$ to features in $\mathscr{F}$. One of the earliest work on zero-shot learning along this line [108] employed wikipedia as a source of semantic concepts used to describe classes. In recent years, embedding spaces learned by deep neural networks are used as semantic spaces.

While some kind of semantics – a representation capable of capturing meaning – is produced in zero-shot learning, there is no such thing in the current approaches to one-shot or few-shot learning. Here, the idea is that at least one example has been seen for each class. This or these example(s) act as representatives of the class. Then, a new example is assigned to the class of the nearest representative. Of course, it all comes down to the definition of "nearest". In this type of learning, the training set, which needs not to be as large as in supervised training, is used in order to learn what is similar or dissimilar. *Contrastive learning* methods in neural networks are one way of realizing this. In this respect, few-shot learning is not very different from case-based reasoning.

5.3 Case-Based Reasoning, Analogical Reasoning and Transfer Learning

Case-based reasoning (CBR for short), e.g., [2] is a form of reasoning that exploits data (rather than knowledge) under the form of cases, often viewed as pairs ⟨problem, solution⟩. When one seeks for potential solution(s) to a new problem, one looks for previous solutions to similar problems in the repertory of cases, and then adapts them (if necessary) to the new problem.

Case-based reasoning, especially when similarity is a matter of degree, thus appears to be close to k-NN methods and instance-based learning [176, 333]. The k-NN method is a prototypical example of transduction, i.e., the class of a new piece of data is predicted on the basis of previously observed data, without any attempt at inducing a generic model for the observed data. The term transduction was coined in [257], but the idea dates back to Bertrand Russell [539].

Another example of transduction is *analogical proportion-based learning*. Analogical proportions are statements of the form "a is to b as c is to d", often denoted by $a : b :: c : d$, which express the claim that "a differs from b as c differs from d and b differs from a as d differs from c". This statement can be encoded into a Boolean logical expression [452, 503] which is true only for the 6 following assignments $(0,0,0,0)$, $(1,1,1,1)$, $(1,0,1,0)$, $(0,1,0,1)$, $(1,1,0,0)$, and $(0,0,1,1)$ for (a,b,c,d). Note that they are also compatible with the arithmetic proportion definition $a-b=c-d$, where $a-b \in \{-1,0,1\}$, which is not a Boolean expression. Boolean Analogical proportions straightforwardly extend to vectors of attributes values such as $\vec{a}=(a_1,...,a_n)$, by stating $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ iff $\forall i \in [1,n],\ a_i : b_i :: c_i : d_i$. The basic analogical inference pattern [588], is then

$$\frac{\forall i \in \{1,...,p\},\ \ a_i : b_i :: c_i : d_i \text{ holds}}{\forall j \in \{p+1,...,n\},\ \ a_j : b_j :: c_j : d_j \text{ holds}}$$

Thus analogical reasoning amounts to finding completely informed triplets $(\vec{a},\vec{b},\vec{c})$ appropriate for inferring the missing value(s) in $\vec{d}$. When there exist several suitable triplets, possibly leading to distinct conclusions, one may use a majority vote for concluding. However it is advisable to use a subset of triplets based on "competent" pairs [83]. This inference method extends to analogical proportions between numerical values, and the analogical proportion becomes graded [216]. It has been successfully applied, for Boolean, nominal or numerical attributes, to classification [451, 85] (then the class $cl(\vec{x})$ (viewed as a nominal attribute) is the unique solution, when it exists, such as $cl(\vec{a}) : cl(\vec{b}) :: cl(\vec{c}) : cl(\vec{x})$ holds). Besides, it has been theoretically established that analogical classifiers *always* yield exact prediction for Boolean affine functions (which includes x-or functions), and only for them [131]. Good results can still be obtained in other cases [132]. *In this issue [84] provides a thorough analysis and experiments regarding the role in analogical inference (applied to classification) played by analogical proportions where the four items involved have the same value for some features.*

It has been also applied to case-based reasoning [411] and to preference learning [232, 81, 82]. The general idea is that a preference between two items can be predicted if some analogical proportions hold that link their descriptions with the descriptions of other items for which preference relations are known. Lastly, analogical inequalities [504] of the form "a is to b at least as much as c is to d" might be useful for describing relations between features in images, as in [394].

The idea of *transfer learning*, which may be viewed as a kind of analogical reasoning performed at the meta level, is to take advantage of what has been learnt on a source domain in order to improve the learning process in a target domain related to the source domain. When studying a new problem or a new domain, it is natural to try to identify a related, better mastered, problem or domain from which, hopefully, some useful information can be called upon for help. The emerging area of transfer learning is concerned with finding methods to transfer useful knowledge from a known source domain to a less known target domain.

The easiest and most studied problem is encountered in supervised learning. There, it is supposed that a decision function has been learned in the source domain and that a limited amount of training data is available in the target domain. For instance, suppose we have learnt a decision function that is able to recognize poppy fields in satellite images. Then the question is: could we use this in order to learn to recognize cancerous cells in biopsies rather than starting anew on this problem or when few labelled data are available in the biological domain?

This type of transfer problem has witnessed a spectacular rise of interest in recent years thanks both to the big data area that makes lots of data available in some domains, and to the onset of deep neural networks. In deep neural networks, the first layers of neuron elaborate on the raw input descriptions by selecting relevant descriptors, while the last layers learn a decision function using these descriptors. Nowadays, most transfer learning methods rely on the idea of transferring the first layers when learning a new neural network on the target training data, fine tuning only the last layers [660, 613, 519]. The underlying motivation is that the descriptors are useful in both the source and target domains and what is specific is the decision function built upon these descriptors. But it could be defended just as well that the decision function is what is essential in both domains while the ML part should concentrate on learning a representation appropriate for the target task. This has been achieved with success in various tasks [126]. When prior knowledge is available in the form of ontologies and some form of logic, it is tempting to use it in order to better select and express what to transfer. For example, in [396], authors propose a framework based on description logic to enhance transfer learning by incorporating semantics and reasoning capabilities to decide when and what to transfer.

Another interesting line of work is related to the study of causality. Judea Pearl uses the term "transportability" instead of transfer learning, but the fundamental issues are the same. Together with colleagues, they have proposed ways of knowing if and what could be transferred from one domain to another [497]. The principles rely on descriptions of the domains using causal diagrams. Thanks to the "do-calculus", formal rules can be used in order to identify what can be used from a source domain to help solve questions in the target domain. One foremost assumption is that causality relationships capture deep knowledge about domains and are somewhat preserved between different situations. For instance, proverbs in natural language are a way of encapsulating such deep causality relationships and their attractiveness comes from their usefulness in many domains or situations, when properly translated.

One central question is how to control what should be transferred. A common assumption is that transfer learning should involve a minimal amount of change of the source domain knowledge in order for it to be used in the target domain. Several ways of measuring this "amount of change" have been put forward (see for instance [134, 569]), but much work remains to be done before a satisfying theory is obtained.

With the recent emergence of Large Language Models, one could be tempted to advocate that the problem of transfer learning is solved. These systems are typically trained on vast amounts of text data from diverse sources, covering a wide range of subjects and domains. And, at least in the field of Natural Language Processing, they are able to answer queries about seemingly any topic. They might thus appear as universal agents.

However, even if these systems are based on the processing of immensely vast databases, they are still limited and have to be adapted for each application domain (e.g. finance, medicine, mathematics, ...). What's more, they require extremely costly learning each time they are updated or adapted to a new domain, which partly defeats the purpose of transfer learning, which aims to make learning for a new task more economical.

Finally, transfer learning can be detrimental. Negative transfer can occur in human learning, when you've been led down the wrong path, and it can also happen in machine learning, and certainly with Large Language Models as well. What are the ingredients for predicting the performance to be expected from transfer learning? Performance of the source hypothesis on the source domain? Proximity (how to measure it?) of the source and target domains? Size of the target training set? Other dimensions? This is still a great unknown and a fruitful area of research given the growing importance of "long life learning".

In this issue [129] discusses and revisit some commonly held beliefs about assumptions needed for transfer learning to work. In particular, it questions the need to have a good hypothesis on the source space and demonstrates that the relation between the source and target space, even when those are different, can matter more than the quality of the source hypothesis. It thus clarifies some of the conditions under which positive transfer learning can occur.

6 Learning for knowledge acquisition

This section focuses on works whose aims are to extract knowledge from data in the form of if-then rules, constraints, or preference relations. We have already encountered rules in logic programming settings in Section 3 when presenting ILP, Statistical Relation Learning and Neural Symbolic AI. In this section, we briefly survey formal concept analysis and rough sets that offer other frameworks for the extraction of rules. Then, we discuss various kinds of if-then rules and their learning, ranging from fuzzy rules, default rules to threshold rules, which respectively take into account the ill-definition of the range of applicability of rules, the presence of exceptions in the data, or express acceptance or rejection in terms of thresholds reached or not reached. Lastly, subsection 6.3 surveys works on constraint acquisition, and subsection 6.4 reports on preference learning and recommender systems. However, some important, well-known formats of knowledge representation, such as Bayesian networks, or other graphical models are not presented in this survey.

In this special issue, [114] presents recent advances on learning and inference using Bayesian networks:

the compilation of Bayesian networks into arithmetic circuits in the form of tensor graphs and the exploitation of partially known functional dependencies leads to significant performance improvements; that paper also shows that one can significantly reduce the reliance on data and improve robustness if one complements data with knowledge, and shows that one can sometimes recover from modelling errors by using the more expressive Testing Bayesian networks.

6.1 Formal concept analysis and rough sets

Formal concept analysis and rough sets are two mathematical frameworks offering original tools for extracting rules from data, which have emerged over the last forty years and developed within dynamic autonomous communities on the fringes of mainstream AI.

6.1.1 Classical rules and formal concept analysis

Formal Concept Analysis (FCA) goes back to the 1980's [259, 244] and can be related to Artificial Intelligence since it offers a formalization of the notion of concept. It is a good example of a setting that stands in between KRR and ML concerns. Moreover recent years have witnessed a renewed interest in FCA with the emergence of Pattern Mining. FCA offers a mathematical framework based on the duality between a set of objects and a set of descriptors or attributes. A set of attributes is often called an itemset in the data mining literature. The basic setting starts from a formal context which is a relation linking objects and (Boolean) attributes. Thus a formal context constitutes a simple repository of Boolean data. A concept is then formalized as a pair composed of a set of attributes and a set of objects representing the intention and the extension of the concept respectively; it has the property that these objects and only them satisfy the set of attributes and this set of attributes refers to these objects and only them. Such a set of attributes is called a closed pattern, an intent, or a closed itemset. More precisely, two operators, forming a Galois connection, respectively associate, to a subset of objects, their common descriptors, and, to a subset of attributes, the subset of objects that satisfy them. In an equivalent way, a pair made of a set of objects and a set of attributes is a formal concept if and only if their Cartesian product forms a maximal rectangle for set inclusion in the formal context. The set of concepts forms a complete lattice. It a possible to define an equivalence relation between itemsets (two sets of attributes are equivalent if they share the same closure). From closed itemsets, it becomes possible to extract, from a formal context, all the attribute implications relating subsets of attributes, of the form “if objects satisfy a set of attributes then they satisfy another set of attributes”. See [289, 259, 48].

Two extensions are especially worth mentioning. One uses fuzzy contexts, where the links between objects and attributes are a matter of degree [52]. Many-valued contexts may be useful for handling numerical attributes [450]. Another extension allows for structured or logical descriptors using so-called pattern structures [258, 243, 28]. Besides, operators other than the ones defining formal concepts make sense in formal concept analysis, for instance, to characterize independent subcontexts [211].

Association rules [9, 297] describe relations between attributes together with support degrees (reflecting the frequency of objects satisfying the condition parts) and confidence degrees (conditional probabilities). Such rules generalise attribute implications, for which the confidence degree is 1 (and that could be supported by no object, i.e., have zero support). Association rules can be extracted from more general data sets than formal contexts: The AMIE system [256] is a well-known example of association rule learner mining an RDF knowledge base in a description logic.

6.1.2 Rough sets and decision rules

Rough sets [493, 118] are imprecise intensional descriptions of sets of objects by means of attribute values in a data base. The idea is that due to the limited number of attributes, objects may be indiscernible in the sense that they have the same description. Indiscernibility is an equivalence relation on the set of objects of the database, determined by a set A of attributes. A subset S of objects in extension can generally only be described by means of upper and lower approximations, each of them being subsets precisely expressed as union of equivalence classes. The lower approximation contains objects certainly belonging to S, and the upper approximation contains objects possibly belonging to S.

Such approximations can be useful in data mining [578, 284], where we face the problem of inconsistent data. As emphasised in [282], "Using rough set theory, conflicting cases are not removed from the data set. Instead, concepts are approximated by new sets called lower and upper approximations."

Based on the lower approximation, rules can be induced that need only cover non-conflicting examples of a class or concept, and must exclude all other examples. For the upper approximation, one looks for rules that cover every example of the class / concept, be it conflicting or not (and must exclude all examples that are not equivalent to any example of the class / concept).

More precisely, consider, in rough set terminology, a decision table $\mathscr{D}$ where there are two kinds of attributes: features $a \in A$ and decisions $d \in D$, which corresponds to a set of examples: every line in the table corresponds to an example, it describes an object (the values for the attributes in A), and an associated decision (or class) - the values for the attributes in D. Each object x described in the decision table can be associated to a decision rule of the form $\wedge_{a\in A} a(x) \Rightarrow \wedge_{d\in D} d(x)$. A decision table can be inconsistent, in the sense that there may be at least two objects x and x', which have the same description in terms of feature attributes $a \in A$ but lead to different decisions.

In the case of inconsistent data, decision rules induced by the decision tables are certainty rules and possibility rules. Consider the set of objects $D(w) = \{x : d(x) = w_d, d \in D\}$. $D(w)$ is the set of objects associated with decision / class w (described by the values $w_d, d \in D$). Let $D_*(w)$ (resp. $D^*(w)$) be the lower (resp. upper) approximation of $D(w)$ induced by attributes $a \in A$. $D_*(w)$ is the set of objects which are necessarily in class w; $D^*(w)$ contains $D_*(w)$ and objects in w for which there exists at least one identical object (in the sense of A) in another class.

This leads to the search for certainty rules, understood as the inclusion $\{x : a(x) = v_a, a \in T\} \subseteq D_*(w)$, and possibility rules, understood as the inclusion $\{x : a(x) = v_a, a \in T\} \subseteq D^*(w)$, for some sets of attributes T. Ideally minimal sets of antecedent attributes are computed. Rough-set-based rule induction systems have been proposed quite early, as for instance the LERS system (see [281] for a bibliography), and its probabilistic extension [283] . Besides, there exist bridges between rough set and formal concept analysis approaches [657] as well as mathematical morphology [29] and Dempster-Shafer theory [101].

6.2 Learning other kinds of if-then rules

Knowledge representation by if-then rules is a format whose importance was early acknowledged in the history of AI, with the advent of rule-based expert systems. Their modelling has raised the question of the adequacy of classical logic for representing them, especially in case of uncertainty where conditioning is often preferred to material implication. Moreover, the need for rules tolerating exceptions, or expressing gradualness on their range of applicability, such as default rules and fuzzy rules has led KRR to develop representation tools beyond classical logic. Those logical formalisms, classical and non-classical, can be used as target languages for learning, understood as knowledge acquisition from data.

In this issue [215] emphasizes that possibility theory stands halfway between logical and probabilistic representation frameworks, and that while qualitative possibility theory is totally compatible with classical logic, quantitative possibility theory can be related to statistics. This suggests that possibility theory may be an interesting setting for interfacing reasoning and learning; it is also a setting of interest for knowledge representation, as recalled in the rest of this subsection.

6.2.1 Default rules

Reasoning in a proper way with default rules (i.e., having potential exceptions) was a challenging task for AI during three decades [90]. Then a natural question is: can rules having exceptions, extracted from data, be processed by a nonmonotonic inference system yielding new default rules? How can we insure that these new rules are still agreeing with the data? The problem is then to extract genuine default rules that hold in a Boolean database. It does not just amount to mining association rules with a sufficiently high confidence level. We have to guarantee that any new default rule that is deducible from the set of extracted default rules is indeed valid with respect to the database. To this end, we need a probabilistic semantics for nonmonotonic inference. It has been shown [57] that default rules of the form "if p then generally q", denoted by $p \rightsquigarrow q$, where $\rightsquigarrow$ obey the postulates of preferential inference [379], have both

1. a possibilistic semantics expressed by the constraint $\Pi(p \wedge q) > \Pi(p \wedge \neg q)$, for any max-decomposable possibility measure Π (a set-function such that $\Pi(p \vee q) = \max(\Pi(p), \Pi(q))$),

2. a probabilistic semantics expressed by the constraint $Prob(p \wedge q) > Prob(p \wedge \neg q)$ for any *big-stepped probability* $Prob$. This is a very special kind of probability such that if $p_1 > p_2 > ... > p_{n-1} \geq p_n$ (where p_i is the probability of one of the n possible worlds), the following inequalities hold $\forall i = 1, n-1,\ p_i > \Sigma_{j=i,n}\, p_j$.

Then, one can safely infer a new default $p \rightsquigarrow q$ from a set of defaults $\Delta = \{p_k \rightsquigarrow q_k | k = 1, K\}$ if and only if the constraints modeling Δ entail the constraints modeling $p \rightsquigarrow q$. Thus, extracting defaults amounts to looking for big-stepped probabilities, by clustering lines describing items in Boolean tables, so as to find default rules, see [55] for details. Then the rules discovered are genuine default rules that can be reused in a nonmonotonic inference system, and can be encoded in possibilistic logic (assuming rational monotony for the inference relation).

It may be also beneficial to rank-order a set of rules expressed in the setting of classical logic in order to handle exceptions in agreement with nonmonotonic reasoning. This is what has been proposed in [561] where a formalization of inductive logic programming (ILP) in first-order possibilistic logic allows us to handle exceptions by means of prioritized rules. The possibilistic formalization provides a sound encoding of non-monotonic reasoning that copes with rules with exceptions and prevents an example from being classified in more than one class.

Possibilistic logic [207] is also a basic logic for handling epistemic uncertainty. It has been established that any set of Markov logic formulas [523] can be exactly translated into possibilistic logic formulas [383, 218], thus providing an interesting bridge between KRR and ML concerns. Taking lessons from this parallel with Markov logic, the learning of default rules encoded as possibilistic logic rules has been futther developed in [384, 385]. However, such a possibilistic encoding may lead to an inference mode that it is not enough cautious with respect to the handling of exceptions; see [58] for a detailed discussion of more or less cautious inference modes in default reasoning. In [590], the authors propose an approach called STRiKE (for STRatified K-Entailment) where k-entailment limits the extent to which erroneous inferences can propagate, by restricting to k the number of constants involved.

Let us also mention two other approaches, also motivated by knowledge completion issues. [255] prefer adding exceptions into the bodies of rules in order to avoid that the prediction of facts by rules introduces errors. In the AnyBURL system (standing for Anytime Bottom Up Rule Learning) [449], rules ordered by their confidence levels are seen as default rules and the approach uses reinforcement learning to find more valuable rules earlier in the search process.

On a quite different basis, let us also mention an attempt at relating non-monotonic inference and neural nets [43, 261].

6.2.2 Possibilistic handling of uncertain rules

A possibilistic handling of a rule-based system was introduced in the 80's [238, 240] and recently revisited in [213]. In this framework, the uncertainty of a rule "if p then q" is handled by a matrix calculus for conditional possibilities, based on max-min composition:

$$\begin{bmatrix} \pi(q) \\ \pi(\neg q) \end{bmatrix} = \begin{bmatrix} \pi(q \mid p) & \pi(q \mid \neg p) \\ \pi(\neg q \mid p) & \pi(\neg q \mid \neg p) \end{bmatrix} \Box^{\max}_{\min} \begin{bmatrix} \pi(p) \\ \pi(\neg p) \end{bmatrix}$$

where the matrix product $\Box^{\max}_{\min}$ uses min as the product and max as the addition. In this matrix calculus, the conditional possibility distributions in the uncertainty propagation matrix obey a qualitative form of conditioning. The max−min composition governing this matrix calculus and the normalization conditions $\max(\pi(p), \pi(\neg p)) = 1$, $\max(\pi(q \mid p), \pi(\neg q \mid p)) = 1$ and $\max(\pi(q \mid \neg p), \pi(\neg q \mid \neg p)) = 1$, ensure that the possibility degrees of the conclusion q are normalized as well. An uncertainty propagation matrix of the form $\begin{bmatrix} 1 & s \\ r & 1 \end{bmatrix}$ where $s, r \in [0, 1]$ are called the rule parameters, represents the rule "if p then q" with certainty $1 - r$ and the rule "if $\neg p$ then $\neg q$" with certainty $1 - s$.

The inference mechanism of a possibilistic rule-based system relies on this matrix calculus. Let a set of n parallel uncertain rules be of the form ($i = 1, \ldots, n$): "if $a_i^1(x)$ is P_i^1 and ... and $a_i^k(x)$ is P_i^k then $b_i(x)$ is Q_i"

relating variables pertaining to the attribute values of some item x, and where the P_i^j's and Q_i are classical subsets of their corresponding attribute domains. In [213], the authors revisited the work of [240] and showed, on an example of a possibilistic rule-based system, that the output possibility distribution obtained by inference can be computed as a min−max product of a matrix containing the rule parameters and a vector whose coefficients are the possibility degrees of the rule premises. The general case for n parallel rules is elaborated in [37]: a min−max matrix relation $O = M \Box^{\min}_{\max} I$ is established, where the matrix containing the rule parameters is constructed by induction on n, the vector I contains the possibility degrees of the rule premises, and the output vector O describes the output possibility distribution on a partition of the output attribute set. In the case of a cascade, i.e. a possibilistic rule-based system composed of two chained sets of parallel possibilistic rules, it is shown in [37] that an input-output relation between the two matrix relations associated to each set of parallel possibilistic rules can be established. The resulting cascade of matrix relations has a structural resemblance to a min-max neural network [37].

These developments lead to two main questions : how to explain the inference results of possibilistic rule-based systems (studied in [240] and subsequently extended in [38, 34]), and how to develop a learning paradigm for determining the rule parameters of possibilistic rule-based systems based on training data [35]. In [35], the author introduced an equation system in order to learn the values of the rule parameters according to a training dataset composed of the possibility degrees of rule premises inferred with the input possibility distributions and a target output possibility distribution. The equation system is composed of a matrix that contains the possibility degrees of rule premises, an output vector which describes a targeted output possibility distribution, and the unknown part is a vector that contains the rule parameters. By applying Sanchez's results on systems of fuzzy relational equations [545] (since standard methods such as gradient descent cannot be easily applied to the min-max composition due to its non-differentiable nature), the author of [35] shows that solving this equation system yields values for the rule parameters that are compatible with the training dataset used. This work was recently extended in [36], where the author showed how to extend the equation system in order to take into account multiple training datasets. The results of [36] on the inconsistency of systems of fuzzy relational equations allows us to obtain approximate solutions of the equation system by modifying as little as possible the right-hand side of the equation system. It can also be used to estimate the quality of a training dataset.

6.2.3 Fuzzy rules

The idea of fuzzy if-then rules was first proposed by Zadeh [662]. They are rules whose conditions and /or conclusions express fuzzy restrictions on the possible values of variables. Reasoning with fuzzy rules is based on a combination / projection mechanism [663] where the fuzzy pieces of information (rules, facts) are conjunctively combined and projected on variables of interest. It generalizes classical logic inference.

Special kinds of fuzzy rules have been used to design fuzzy rule-based controllers in systems engineering: fuzzy rules may specify the fuzzy graph of a control law, as the union of fuzzy granules; once applied to an input it yields a fuzzy output that is usually defuzzified [433]. Other kinds of rules have precise conclusions that are combined on the basis of the degrees of matching between the current situation and the fuzzy condition parts of the rules [592]. In both cases, an interpolation mechanism is at work, implicitly or explicitly [664]. Such fuzzy rule-based controllers are universal approximators [105] and can be used to represent continuous non-linear systems.

There are other kinds of fuzzy rules whose primary goal is not to approximate functions, but rather to offer mathematical representations of various kinds of if-then statements in natural language [209]. This is, for instance, the case of gradual rules, which express statements of the form "the more x is A, the more y is B", where A and B are gradual properties modelled by fuzzy sets [559, 486]. It is possible to quantify the association between condition and consequent parts for such fuzzy rules with respect to a database. See [206] for the proper assessment of confidence and support degrees for fuzzy rules. Fuzzy rough sets have been used for extracting gradual decision rules [277]. The mining of gradual patterns has been studied quite extensively by Laurent *et al.*, see, e.g., [486].

The functional equivalence between a radial basis function-based neural network and a fuzzy inference system has been established under certain conditions [352]. Moreover, fuzzy rules may provide a rule-based interpretation [143, 541] for (simple) neural nets, and neural networks can be used for extracting fuzzy rules from the training data [498, 254].

In the perspective of classification, learning methods for fuzzy decision trees have been devised in [445], in the case of numerical attributes. The use of fuzzy sets to describe associations between data and decision trees may have some interest: extending the types of relations that may be represented, making easier the interpretation of rules in linguistic terms [219], and avoiding unnatural boundaries in the partitioning of the attribute domains.

6.2.4 Threshold rules

Another format of interest is the one of multiple threshold rules, i.e., selection rules of the form "if $x_1 \geq \alpha_1$ and $\cdots$ $x_j \geq \alpha_j$ and $\cdots$ then $y \geq \gamma$" (or deletion rules of the form 'if $x_1 \leq \beta_1$ and $\cdots$ $x_j \leq \beta_j$ and $\cdots$ then $y \geq \delta$"), which are useful in monotone classification / regression problems [278]. An algorithm [74], called VC-DomLEM, based on a rough set approach (where the equivalence relation is replaced by a dominance relation for handling ordered data) provides a mechanism for inducing such rules. Indeed when dealing with data that are made of a collection of pairs $(x^k, y_k), k = 1, ..., N$, where x^k is a tuple $(x_1^k, ..., x_n^k)$ of feature evaluations of item k, and where y is assumed to increase with the x_i's in the broad sense, it is of interest of describing the data with such rules of various lengths. It has been noticed [279, 217] that, once the numerical data are normalized between 0 and 1, rules where all (non trivial) thresholds are equal can be represented by Sugeno integrals (a generalization of weighted min and weighted max, which is a qualitative counterpart of Choquet integrals [275]). Moreover, it has been shown recently [89] that generalized forms of Sugeno integrals are able to describe a global (increasing) function, taking values on a finite linearly ordered scale, under the form of general thresholded rules. Another approach, in the spirit of the version space approach [461], provides a bracketing of an increasing function by means of a pair of Sugeno integrals [506, 505].

6.3 Constraint learning

Constraint Programming (CP) [533] is a declarative framework for solving combinatorial problems. In such a framework, a problem is specified by a set of variables $\mathscr{V}$, a set of domains $\mathscr{D}$, a domain D_i for each variable x_i and a set $\mathscr{C}$ of constraints, where a constraint is put on a subset of variables and specify the values that are allowed. For instance, if x_1 and x_2 are two variables, a constraint $x_1 \neq x_2$ means that in a solution of the problem, the values of x_1 and x_2 must be different. Such a formalization of the combinatorial problem is called a constraint network. Two kinds of problems are considered: satisfaction problems that aim at finding an assignment of the variables of $\mathscr{V}$ satisfying all the constraints and optimization problems that given an optimization criteria aims at finding the best solution satisfying the constraints. Solving a problem in CP is done through two operations: constraint propagation, or filtering, allowing to reduce the possible domains of variables by removing inconsistent values from the variable domains and branching that consists in choosing a variable, splitting its domain and creating branches in the search tree, one for each split. Solvers have been made more efficient by the development of global constraints embedding a set of constraints and for which more powerful filtering algorithms have been defined. For instance a *alldifferent* global constraint expressing that a set of variables must be assigned to different values has been defined and a filtering algorithm has been developed, which is more efficient than considering the conjunction of elementary pairwise different constraints between variables. The efficiency of CP depends on the way a problem is modeled (the choice of the variables and of the constraints) and on the search strategies for branching, making its use difficult for a non expert of CP. A new research domain, called *constraint acquisition* aims at learning a constraint network, given a library of constraints and a set of positive and possibly negative examples of assignments of variables for the target constraint network.

Constraint acquisition is often formulated as a concept learning problem, where the hypothesis space is defined through biases put on the form constraints can take and positive (resp. negative examples) are specified by solutions (resp. non solutions) to the combinatorial problem. The system Conacq introduced in [124] relies on a version space algorithm [463], formalized as a SAT-problem in [68], whereas [391] proposes a framework based on Inductive Logic Programming. [516] proposes an interesting study on the relations between constraint network learning and inductive logic programming (ILP), formalizing different works in a ILP setting. It emphasizes the specificities of constraint learning, in particular the fact that the number of positive examples is usually small and that the number of constraints can be large with many syntactic variants.

In this issue, [508] propose a novel, statistical approach based on sequential analysis that is fast, can handle large biases, and can accurately learn constraints from noisy data.

These approaches suffer for several drawbacks: all positive and negative examples must be given at the beginning of the learning process, which might be difficult for an expert and a bias on the constraint language must be specified (e.g. a set of constraints as in [68], a catalog of global constraints as in [50])

[69] introduces active constraint learning, where membership queries that are complete instantiations of variables are generated and presented to the user, who has to classify each query as positive or negative examples. The idea of using queries to learn an unknown concept was first introduced in Machine Learning in [21]. Convergence is reached when (1) a constraint network satisfying all the positive and negative examples is found and (2) all other constraint networks satisfying (1) are equivalent, i.e. have the same set of solutions. As in all active learning systems, the problem is then to generate queries that are sufficiently informative to reduce the number of queries that have to be asked to the user before converging to the expected constraint network. [69] introduces the notion of irredundant queries, i.e. queries that cannot be inferred by the current constraint network and the notion of optimistic queries. A synthesis of the work performed around the system Conacq and its extension to active constraint learning can be found in [71]. Nowadays most works on active learning consider only partial queries and many variants have been introduced to limit the number of queries [67, 26, 608, 65, 606]. Nevertheless the efficiency of the systems depend on the size of the constraint language. Meta algorithms [604, 605] have been proposed to overcome this: the first one tends to call a constraint acquisition system on a growing set of variables, thus starting by a small bias (constraint language), the second one proposes to use a probabilistic classification model to generate more promising queries. Extensions have been proposed: [607] introduces Limited Member Queries for which the user can answer 'yes', 'no' or 'I don't know'; [66] proposes a weighted partial max-sat approach to learn a constraint network over unknown constraint language, given the assumption that the size of the language and the arity of the relations between variables being fixed.

In this section we have focused on hard constraint learning. [517] presents a more general, logical-based view of constraint learning, including soft constraint learning.

6.4 Preferences and recommendation

The importance of the notion of preference seems to have emerged first in economics and decision theory, and research in these fields focused essentially on *utilitarian* models of preferences, where utility function associates a real number with each one of the objects to be ordered. Also in this field, research developed on *preference elicitation*, where some interaction is devised to help a decision maker form / lay bare her preferences, usually over a relatively small set of alternatives, possibly considering multiple objectives.

In contrast, preferences in AI often bear on combinatorial objects, like models of some logical theory to indicate for instance preferences over several goals of an agent; or, more recently, like combinations of interdependent decisions or configurable items of some catalog. Thus, in KRR as well as in ML, the objects to be ordered are generally characterised by a finite number of features, with a domain / set of possible values for each feature. When talking about preferences, the domains tend to be finite ; continuous domains can be discretised.

Preferences are now an important ingredient in AI. It has been an active research topic in several communities, in particular in AI and Operations Research, as shown by e.g. journal special issues [357, 196]), continuing series of workshops (DA2PL,MPREF).

In both KRR and ML, models have evolved from binary ones (classical propositional or first-order logic, binary classification) to richer ones that take into account the need to propose less drastic outputs. One approach has been to add the possibility to *order* possible outputs / decisions. In multi-class classification tasks for instance, one approach [203] is to estimate, given an instance, the *posterior* probability of belonging to each possible class, and predict the class with highest probability. The possibility of learning to "order things" has numerous applications, e.g., in information retrieval, recommender systems. In KRR, the need to be able to order interpretations (rather than just classify them as possible / impossible, given the knowledge at hand) has proved to be an essential modelling paradigm, see, e.g., the success of valued CSPs [549], Bayesian networks [494], possibilistic / fuzzy logics among others.

At the intersection of ML and KRR, the field of "preference learning" has emerged. Furnkranz et al. [252, 253] describe various tasks that can be seen as preference learning, some where the output is a function that orders possible labels for each unseen instance, and some where the output is a function that orders any unseen set of new instances.

Because of the combinatorial nature of the space of objects, research in AI emphasized the need for *compact* models of preferences. Some probabilistic models, like Bayesian networks or Markov random fields, fall in this category, as well as e.g., additive utilities [248] and their generalisations. This focus on combinatorial objects also brought to light one difficulty with the utilitarian model: although it is often easy to compute the utilities or probabilities associated with two objects and compare them on such a basis, it appears to be often NP-hard to find optimal objects from a combinatorial set with numerical representations of preferences. Thus one other contribution of research in KRR is to provide preference representation languages where optimisation is computationally easy, like CP-nets [88]. See e.g. [236, 235] for comparisons of the complexity of queries for some popular preference representation models.

These complex models of preferences have been studied from an ML perspective, both in an elicitation / active learning setting, and in a batch / passive learning setting. One particularity of these compact preference models is that they combine two elements: a structural element, indicating probabilistic or preferential interdependencies between the various features characterizing the objects of interest; and "local" preferences over small combinations of features. It is the structure learning phase which is often demanding, since finding the structure that best fits some data is often a hard combinatorial search problem. In contrast, finding the local preferences once the structure has been chosen is often easy.

The passive learning setting is particularly promising because of the vast dataset available in potential applications of preference learning in some decision aid systems like recommender systems or search engines. The possibility to learn Bayesian networks from data has been a key element for their early success in many applications. Note that in some applications, in particular in the study of biological systems, learning the structure, that is, the interdependencies between features, is interesting; in such applications, "black-box" models like deep neural networks seem less appropriate. This is also the case in decision-support systems where there is a need to *explain* the reasons justifying the computed ordering of possible decisions [49].

At the frontier between learning and reasoning lies what could be named lazy preference learning: given a set of preference statements which do not specify a complete preference relation, one can infer new pairwise comparisons between objects, by assuming some properties the full, unknown preference relation. As a baseline, many settings in the models studied in KRR assume transitivity of preferences, but this alone does not usually induce many new comparisons. A common additional assumption, made by [49], is that the preference relation can be represented with an additive utility function, and that the ordering over the domain of each feature is known. In [637, 638, 639], richer classes of input preference statements are considered, and the assumption is made that the preference relation has some kind of (unknown) lexicographic structure.

At first sight, the task of recommending products, such as restaurants, movies, goods etc. could be considered as a natural application field of preference learning, albeit with a truly impressive scale. For instance, the Amazon platform sells 350 million products to 300 million customers[5] and Youtube has 2.6 billion monthly active users who can potentially watch more than 5 billion videos[6]. But, actually, the fields of preference learning on the one hand, and automatic recommendation on the other, have followed essentially distinct evolutions.

The rise of the web in the nineties started the interest in recommender systems. Earlier recommendations systems were mostly *content-based* [42]. In this approach, items are described with descriptive sets of attributes. Content-based recommender systems try to match users to items that are similar to what they have rated high in the past. Users are described by their profile. A profile may be an explicit set of preferences expressed by the user, say "I like spy stories with lot of actions", or it can be based on the ratings of past seen products. One advantage is that it is easy to introduce new products as long as they are properly described. The recommendations for one user do not depend on the other users. On the other

[5] https://www.bigcommerce.com/blog/amazon-statistics/

[6] https://www.globalmediainsight.com/blog/youtube-users-statistics/

hand, a disadvantage is that content-based recommendations tend to be overspecialized on the user and do not offer a high level of serendipity, that is they often present little surprise to the user.

Because it is not easy to devise before hand the set of relevant attributes for describing products and users, but there exists an enormous amount of interaction history between the usually very large set of users and products, another approach has gained prominence in the first decade of the 21st century: *collaborative-based* recommendations. In this approach, the history of interactions between users and potential items is expressed as an enormous preference matrix where each cell condenses the judgement of one user on one item if this user has expressed explicitly or implicitly an opinion about this item in the past. Obviously, this matrix is extremely sparse, since each user has seen at most several hundreds of items out of possibly millions. The whole trick is therefore to find methods in order to "fill" the missing entries. One common way to fill the empty cells of the preference matrix is to resort to Singular Value Decomposition (SVD) which compresses the matrix through a linear approximation and then re-expands it. One measure of performance commonly used is the Mean-Absolute-Error (MAE). In essence, the ratings of like-minded users to the target one, because they have rated some common items in the past in a similar way, are used in order to make the recommendations. For a given item, a weighted average of the ratings expressed by this peer group is used to evaluate the likely rating that the target user would give to this item.

Another way to compress the information of the preference matrix has been proposed [302], which uses neural networks trained to predict the relevance scores of user-item pairs, therefore enabling to return personalized top-K items for each user.

Recently, Large Language Models (LLMs) have attracted attention in the field of recommending systems. These models trained on massive amounts of data have demonstrated remarkable success in learning what appears as universal representations. In the context of recommendation making, they can leverage the extensive knowledge encoded within them and respond to users' prompts by high quality textual answers and explanations [640].

From these successive phases of the field, it can be seen that after the period of content-based methods, approaches have tended to ignore altogether conceptual descriptions of users and items to rely exclusively on the exploitation of the preference matrix. The recent turn towards using LLMs reintroduces the notions of knowledge representations, albeit in a not straightforward manner. Lot of research remains to be done to better understand these "foundation models", their true potential and their limits.

7 Learning for reasoning

This topic focuses on situations where learning is exploited so as to help or guide reasoning tasks, as a reciprocal to section 5, where reasoning is exploited to help perform a learning task. We also do not consider here learning for knowledge acquisition, where the results provided by the learning step offer valuable knowledge that can be reused in a larger reasoning system: this has been covered in the previous section.

This section provides a brief overview of the applications of ML in practical Automated Reasoners[7]. With a few exceptions, the subsection emphasizes recent work, published over the last decade. Tightly related work, e.g. inductive logic programming (see Section 3.1) or statistical relational learning [170], is beyond the scope of this subsection. The section is organized in two main parts. First, 7.1 overviews the applications of ML in developing and organizing automated reasoners. Subsection 7.2 covers a number of recent topics at the intersection of automated reasoning and ML.

Recent uses of automated reasoning in learning ML models, improving the robustness of ML models, but also in explaining ML models, are covered in Section 8.

7.1 Machine Learning vs. Automated Reasoners

Until recently, the most common connection between ML and automated reasoning would be to apply the former when devising solutions for the latter. As a result, a wealth of attempts have been made towards

[7] We adopt a common understanding of *Automated Reasoning* as "*The study of automated reasoning helps produce computer programs that allow computers to reason completely, or nearly completely, automatically*" (from https://en.wikipedia.org/wiki/Automated_reasoning).

applying ML in the design of automated reasoners, either for improving existing algorithms or for devising new algorithms, built on top of ML models. Uses of ML can be organized as follows. First, uses of ML for improving specific components of automated reasoners, or for automatic configuration or tuning of automated reasoners. Second, approaches that exploit ML for solving computationally hard decision, search and counting problems, and so offering alternatives to dedicated automated reasoners.

Improving Reasoners. Earlier efforts on exploiting ML in automated reasoners was to improve specific components of reasoners by seeking guidance from some ML model. A wealth of examples exist, including the improvement of restarts in Boolean Satisfiability (SAT) solvers [296], improvement of branching heuristics [249, 280, 408, 407, 409], selection of abstractions for Quantified Boolean Formulas (QBF) solving [353, 397], but also for improving different components of theorem provers for first-order and higher order logics [609, 360, 363, 361, 345, 423, 120]. It should be noted that the performance gains obtained in automated reasoners (e.g. SAT solvers) over the past three decades hinge in good part from changing the original logic formula by learning new clauses [441]. While the learning of new clauses cannot be viewed as the direct result of using machine learning, one can envision applying machine learning for improving the learning of new clauses.

ML has found other uses for improving automated reasoners. A well-known example is the organization of portfolio solvers [650, 335, 337, 376, 447]. Another example is the automatic configuration of solvers, when the number of options available is large [336, 576]. One additional example is the automatic building of automated reasoners using ML [376].

Tackling Computationally Hard Problems. Another line of work has been to develop solutions for solving computationally hard decision and search problems. Recent work showed promise in the use of NNs for solving satisfiable instances of SAT represented in clausal form [555, 556, 554, 627], for solving instances of SAT represented as circuits [14, 15], but also NP-complete problems in general [507]. The most often used approach has been to exploit variants of Graph Neural Networks (GCNs) [548], including Message Passing Neural Networks (MPNNs) [266]. There has also been recent work on solving CSPs [647] using convolutional NNs. Furthermore, there have been proposals for learning to solve SMT [44], combinatorial optimization problems [374, 51, 405, 61], planning problems [195], but also well-known specific cases of NP-complete decision problems, e.g. Sudoku [489] and TSP [619].

Efforts for tackling computationally harder problems have also been reported, including QBF [656], ontological reasoning [316], probabilistic logic programming [434], inference in probabilistic graphical models [659] and theorem proving for first order [362, 325, 46, 488, 324] and higher order logics [636, 626, 359, 655].

7.2 More on Learning vs. Reasoning

The previous two subsections summarize recent efforts on using machine learning for automated reasoning, but also on using automated reasoning for learning, verifying and explaining ML models. This subsection identifies additional lines of research at the intersection of ML and automated reasoning.

Integrating Logic Reasoning in Learning. A large body of work has been concerned with the integration of logic reasoning with ML. One well-known example is neural-symbolic reasoning [260, 161, 64, 491, 72, 436, 155]. See also Subsection 3.2. Examples of applications include program synthesis [673, 672, 491, 94] and neural theorem proving [456]. Other approaches do exist, including deep reasoning networks [112], neural logic machines [198], and abductive learning [142, 677]. An alternative is to embed symbolic knowledge in neural networks [644].

Learning for Inference. One area of work is the use of ML models for learning logic representations, most often rules [526, 527, 654, 229, 228], which can serve for inference or for explaining predictions. See also Subsection 3.2.

Understanding Logic Reasoning. A natural question is whether ML systems understand logical formulas in order to decide entailment or unsatisfiability. There has been recent work on understanding entailment [230, 547], suggesting that this is not always the case, e.g., for convolutional NNs. In a similar fashion, recent work [115] suggests that GNNs mail fail at deciding unsatisfiability.

Synthesis of ML Models. Recent work proposed the use of automated reasoners for the synthesis (i.e., learning) of ML models. Concrete examples include [247, 678, 415]. These approaches differ substantially from approaches for the synthesis of interpretable models, including decision trees and sets and decision lists.

As witnessed by the large bibliography surveyed in this subsection, the quantity, the breadth and the depth of existing work at the intersection between ML and automated reasoning in recent years, provides ample evidence that this body of work is expected to continue to expand at a fast pace in the near future.

8 Model accountability

For accountability and intelligibility purposes, it may be desirable to ensure that a model possess some abilities, or obey to some natural or imposed constraints. This typically corresponds to requiring the model to have additional reasoning features mainly coming in the form of symbolic structures: being readable by a human expert; modelling or discovering causal, non-symmetric relationships; being able to explain in some sense the made predictions. In this section, we are concerned with approaches that increase the model accountability and intelligibility. This concern has been recently emphasized at an institutional level, AI regulation texts, such as the AI Act from the European Parliament mentions that users need to dispose of a "right to explanation", while remaining vague on the definition of the latter.

Data intelligibility may also be increased by means of rules extracted from the data, e.g. gradual rules, getting structured outputs and patterns from data, cluster labelling, ...); this point is outside the scope of this section and has been covered (at least in part) in Section 6.

This section first surveys recent works about explainability and interpretability, before addressing the issues of robustness and fairness in machine learning.

8.1 Explainability and interpretability

This subsection overviews the uses of automated reasoning approaches for verifying, explaining and learning ML models.

Explanations with Abductive Reasoning. In many settings, interpretable models are not often the option of choice, being replaced by so-called black-box models, which include any ML model from which rules explaining predictions are not readily available. [8] Concrete examples include (deep) neural networks (including binarized versions), and boosted trees and random forests, among many other alternatives.

Most existing works on computing explanations resort to so-called *local* explanations. These models are agnostic and heuristic in nature [521, 426, 522]. Recent works [479, 343] revealed that local explanations do not hold globally, i.e., it is often the case that there are points in feature space, for which the local explanation holds, but for which the model's prediction differs. Since 2018, a number of attempts have been reported, which propose rigorous approaches for computing explanations. Concretely, these recent attempts compute so-called *abductive* explanations, where each explanation corresponds to a prime implicant of the discrete function representing the constraint that the ML model predicts the target prediction. A first attempt based on compiling such a function into a tractable representation is reported elsewhere [571]. For such a representation, (shortest) prime implicants can then be extracted in polynomial time. The downside of this approach is that compilation may yield exponential size function representations. Another

[8] The definition of *explanation* is the subject of ongoing debate [454]. We use the intuitive notion of explanation as a IF-THEN rule [521, 426, 522], where some given prediction is made if a number of features values hold true. The importance of reasoning about explanations is illustrated by a growing number of recent surveys [312, 314, 73, 469, 377, 313, 5, 11, 200, 315, 288, 543, 544, 454, 453, 22, 465, 646, 471].

attempt [342] is based on computing explanations of demand, by encoding the instance, the ML model and the prediction into some logic representation. In this case, reasoners such as SMT, ILP or SAT solvers are then used for extracting (shortest) prime implicants.

The progress observed in computing abductive explanations in recent years is summarized in several recent works [439, 437, 147]. More importantly, by relating abductive explanations with adversarial examples, recent work demonstrated that abductive explanations can be computed for fairly complex block-box ML models [327, 641].

In this issue the paper [16] is directly concerned with abductive and logical explanations. it proposes to extend necessary explanation of robust additive models by possible, yet non-necessary steps in order to explain a preference in order to extend the scope of explanation engines. Among other things, this allows the decision maker to scrutinize and confirm or criticize the provided explanations.

Interpretable ML Models. Interpretable ML models are those from which rule-like explanations can be easily produced. For example, decision trees, decision sets (or rule sets) and rule lists are in general deemed interpretable, since one can explain predictions using rules. On area of research is the learning (or synthesis) of interpretable ML models using automated reasoning approaches. There have been continued efforts at learning decision trees [484, 483, 70, 485, 616, 477, 617, 614, 321, 8, 173], decision sets [390, 344, 430, 265] and rule lists [18, 19]. Examples of reasoners used include SAT, CP, and ILP solvers, but dedicated complete methods based on branch and bound search have also been considered. Despite a recent explosion of works on black-box ML models, there exist arguments for the use of interpretable models [536].

In this issue [304] intends to learn a set of easy-to-interpret models, namely scoring rules (notably used by physicians), so that the decision model can be adapted to the context as well as to the information at hand.

Despite ever-increasing efforts towards the learning of interpretable models, it has been shown that these so-called interpretable models must be explained, i.e. explanations for predictions are not trivially obtained by manual inspection of such models [346, 347, 440], and so explanations (see the following section) must be computed.

Explanations vs. Adversarial Examples. In recent years, different works realized the existence of some connection between adversarial examples (AE's) and explanations (XP's) [420, 593, 532, 597, 649, 490, 107]. Nevertheless, a theoretical connection between AE's and XP's has been elusive. Recent work [342] showed that adversarial examples can be computed from the set of explanations for some prediction. Furthermore, this work introduced the concept of counterexample (CEx) to some prediction, and identified a minimal hitting set relationship between XP's and CEx's, i.e., XP's are minimal hitting sets of CEx's and vice-versa.

Later work established minimal hitting-set duality between the sets of abductive and contrastive explanations [340]. In practice, this duality relationship serves for the enumeration of both abductive and contrastive explanations.

In this issue [666] studies the difficulty of performing counter-factual queries that explicitly account for causal structures. In particular, it shows that sets of probabilities (a.k.a. credal sets) are well-adapted to deal with such queries, but that their exact resolution is NP-hard, offering a an efficient heuristic to bypass this computational bottleneck. It therefore addresses two important aspects of accountability that are counter-factuals and causality, this last aspects being also important from a modelling perspective. Counter-factuals are also strongly connected to explanations as well as adversarial examples.

8.2 Robust Machine Learning

Concerns about the behavior of neural networks can be traced at least to the mid 90s and early 00s [634, 667, 552]. Additional early work on ensuring safety of neural networks also involved SAT solvers [511]. More recently, efforts on the verification of neural networks have focused on the avoidance of so-called adversarial examples.

Adversarial examples [591], already briefly mentioned in Subsection 2.4, illustrate the brittleness of ML models. In recent years, a number of unsettling examples served to raise concerns on the fragility neural networks can be in practice [30, 231, 106, 246, 303]. Among other alternative approaches, automated reasoners have been applied to ensure the robustness of ML models, emphasizing neural networks. A well-known line of work focuses on the avoidance of adversarial examples for neural networks using ReLU units [475] and proposes Reluplex, an SMT-specific dedicated reasoning engine for implementing reasoning with ReLU units [369, 273, 370]. Another recent line of work addresses binarized neural networks [329] and develops a propositional encoding for assessing the robustness of BNNs [478, 476, 400]. Additional lines of work have been reported in the last two years [326, 535, 103, 386, 263, 460, 573, 575, 574, 33, 221, 97, 220, 512, 222, 624, 96, 632, 95, 564, 201, 45]. In the case of complex neural networks, recent advances in robustness tools have been assessed in a dedicated competition [91].

8.3 Fairness

The problem of fairness in machine learning algorithms arises when such algorithms are applied to critical tasks that impact people. By design, machine learning algorithms aim at identifying biases in the data and generalizing them for decision-making tasks. However, the use of certain features, such as gender, political or religious opinions, ..., is prohibited by law or may not align with ethical considerations [242, 202].

Fairness usually addresses situations where an algorithm exhibits different behavior for two distinct subgroups of the population, although these subgroups should not influence its outcome. This situation is often modeled as follows: the algorithm should aim at forecasting a variable Y based on observations X. Fairness is then defined with respect to a protected variable, called a protected attribute, S, which represents membership in each population subgroup. An algorithm is considered fair if its predictions do not depend too much on S. The question of fairness in machine learning is first to evaluate if a model's decision is influenced by a protected variable S and, in such cases, correct the model to remove this dependency [223, 669].

In this issue, [75] propose an exact learning algorithm to extract a Horn theory that corresponds in some sense to a given learnt model: this Horn theory can be used to explore existing biases in the model; they apply it in particular to probe occupational gender biases in BERT-based language models.

The bias in the data may come from many sources, such as existing societal biases or biased sampling of the data [251]. It is important to note that simply removing the protected variable from the dataset is generally not sufficient since other variables may be correlated with S [152]. Moreover, in some situations, such as machine learning with images, removing a feature is not straightforward. There are two major approaches to correcting the fairness of a model. The first approach consists of repairing the dataset by removing the dependence with respect to the protected variable with minimal alteration of the dataset [242, 251]. Another way to achieve fairness is to constrain the model to make fair decisions [665].

9 Handling imperfect data and information in learning methods

The idea of combining KRR notions with ML tools is not new. In KRR it is usual to deal with imperfect information, especially incomplete one: it is very rare that a knowledge base is complete. This situation contrasts with statistical and machine learning, that assumes by default that data, predictions are either precise or random. It of course does not mean that statistics and machine learning never deal with missing information, as we shall discuss in the next sections.

We will discuss the general problems on which we focus in Section 9.1, and will then make a focus on various methods mixing KR and ML components.

9.1 Uncertainty in ML: in the data and in the model

We will focus on two aspects in which cross-fertilisation of ML with KRR could be envisaged: uncertainty in the data and uncertainty in the models/predictions. Note that we will not deal with imperfections due to an adversarial modification of the data, a topic partially covered in Section 8.2.

9.1.1 Learning under uncertain and coarse data

In general, learning methods assume the data to be complete, typically in the form of examples modelled by (tuples of) precise values (in the unsupervised case) or precise input/output pairs (in the supervised case). There are however various situations where data can be expected to be uncertain, such as when they are provided by human annotators in classification, or measured by low-quality sensors, or yet even missing, such as when sensors have failed or when only few examples could be labelled. An important remark is that the uncertainty attached to a particular piece of data can hardly be said to be of objective nature (representing frequency) as it has a unique value, and this even if this uncertainty is due to a random process.

While the case of missing (fully imprecise) data is rather well-explored in the statistical [418] and learning [109] literature, the general case of uncertain data, where this uncertainty can be modelled using different representation tools of the literature, largely remains to be explored. In general, we can distinguish between two strategies:

- The first one intends to extend existing methods for precise data to the handling of uncertain data, while retrieving a precise model from them. The most notable approaches consist in either extending the likelihood principle to uncertain data (e.g., [178] for evidential data, or [136] for coarse data), or to provide a precise loss function defined over partial data and then using it to estimate the empirical risk, see for instance [330, 133, 121, 332]. Such approaches are sometimes based on specific assumptions, usually hard to check, about the process that makes data uncertain or partial. Some other approaches such as the evidential likelihood approach outlined in [178] do not start from such assumptions, and simply propose a generic way to deal with uncertain data. We can also mention transductive methods such as the evidential K-nearest neighbour (K-NN) rule [176, 189, 186], which allows one to handle partial (or "soft") class labels without having to learn a model. *In this issue, [100] studies the complexity of applying generalized risk minimisation (GRM) to uncertain supervised data when uncertainty is described by fuzzy sets. It shows in particular that obtaining good models with theoretical guarantees using GRM may be computationally challenging, and proposes an alternative learning using randomisation and ensembling that is efficient and provides good empirical performances*;

- The second approach, much less explored, intends to make no assumptions at all about the underlying process making the data uncertain, and considers building the set of all possible models consistent with the data. Again, we can find proposals that extend probability-based approaches [163], as well as loss-based ones [138]. The main criticisms one could address to such approaches is that they are computationally very challenging and also sometimes too conservative, as adding imprecise/uncertain instances will systematically enlarge the models. Moreover they do not yield a single predictive model, making the prediction step potentially difficult, even if more robust. In theory, such criticisms can be mitigated by trying to find a compromise between being fully skeptical and picking an inductive bias leading to a single, precise model, as sometimes done when learning imprecise models [24]. Although such middle ground approaches are still very rare for uncertain data, one can find some attempt, e.g., in preference or voting modelling [380, 233].

The problem of handling partial and uncertain data is certainly widely recognised in the different fields of artificial intelligence, be it KRR or ML. One remark is that mainstream ML has, so far, almost exclusively focused on providing computationally efficient learning procedures adapted to imprecise data given in the form of sets, as well as the associated assumptions under which such a learning procedure may work [419]. While there are proposals around that envisage the handling of more complex form of uncertain data than just sets, such approaches remain marginal, at least for two major reasons:

- More complex uncertainty models require more efforts at the data collection step, and the benefits of such an approach (compared to set-valued data or noisy precise data) do not always justify the additional efforts. However, there are applications in which the modelling of data uncertainty does improve the performance of classification tasks [117, 514], or allow to robustify the results of the learning algorithm [413, 414]. Another possibility could be that those data are themselves predicted by an uncertain model, then used in further learning procedures, as for example already done in stacking [225] or self-supervised procedures [412, 528], in which case increased performances have also been observed;

- Using more complex representations may involve a higher computational cost, and the potential gain of using such representations is not always worth the try. However, some specific methods extended to the belief function setting such as the EM algorithm [178] or the K-NN rule [186] , make it possible to handle uncertain data without additional cost. Similarly, some adaptations of the cost functions together with a suitable choice of uncertainty representation lead to differentiable learning procedure presenting the same computational complexity as standard learning [413].

9.1.2 Uncertainty in the prediction model

Another step of the learning process where uncertainty can play an important role is in the characterisation of the model or its output values. In the following, we will limit ourselves to the supervised setting where we try to learn a (predictive) function $f : \mathscr{X} \to \mathscr{Y}$ linking an input observation $x \in \mathscr{X}$ to an output (prediction) $y \in \mathscr{Y}$. Assessing the confidence one has in a prediction can be important in sensitive applications. This can be done in different ways:

- By directly impacting the model f itself, for instance associating to every instance x not a deterministic prediction $f(x)$, but an uncertain output on the domain $\mathscr{Y}$. The most common type of output is of course probability distributions, but other solutions such as possibility distributions, belief functions or convex sets of probabilities are possible;

- Allowing the prediction to become imprecise, the main idea behind such a strategy is to have weaker yet more reliable predictions. In the classical setting, this is usually done by an adequate replacement of the loss function [295, 276], yet recent approaches take a different road. For instance, imprecise probabilistic approaches consider sets of models combined with a skeptic inference (also a typical approach in KR), where a prediction is rejected if it is so for every possible model [125]. Approaches to quantify statistical predictions in the belief function framework are described in [364, 652, 187] and, more recently, in the epistemic random fuzzy set framework in [181, 183]. Conformal prediction and approaches [566, 20], rooted in the old idea of order and non-parametric statistics, is another approach that can be plugged in to any model output to obtain set-valued predictions, and that is gaining traction in machine learning, due to its simplicity, flexibility and theoretical justification.

If such approaches are relatively well characterised for the simpler cases of multi-class classification, their extension to more complex settings, such as multi-label or ranking learning problems that involve combinatorial spaces, remain largely unexplored, with only a few contributions (see [116, 25, 233, 10, 482, 188] for a few ones). It is quite possible that classical AI tools such as SAT or CSP solvers could help deal with such combinatorial spaces.

In this issue [468] studies probabilistic circuits, that are emerging as an efficient tool to integrate knowledge in learning techniques, as well as to derive generative models with a clear probabilistic semantic. They are also computationally attractive, as many queries can be performed in polynomial time. The authors of this paper study their robust counter-part, where probabilities are allowed to become imprecise. It shows that in such a situation, performing queries becomes non-polynomial in many cases, but the paper proposes efficient and accurate heuristics to solve this issue.

9.2 Dempster-Shafer Reasoning and Generalized Logistic Regression Classifiers

The theory of belief functions originates from Dempster's seminal work [174] who proposed, at the end of the 1960's, a method of statistical inference that extends both Fisher's fiducial inference and Bayesian inference, generating imprecise probabilities. In a landmark book, Shafer [565] reconsidered Dempster's lower and upper probabilities and extended their domain of application to the representation of subjective belief. Shafer showed that it could be proposed as a general language to express "probability judgements" (or degrees of belief) induced by items of evidence. This new theory rapidly became popular in Artificial Intelligence where it was named "Dempster-Shafer (DS) theory", evidence theory, or the theory of belief functions [184]. DS theory can be considered from different perspectives:

- A belief function can be defined axiomatically as a Choquet monotone capacity of infinite order [565].
- Belief functions are intimately related to the theory of random sets: any random set induces a belief function and, conversely, any belief function can be seen as being induced by a random set [481]. However, it is a random *disjunctive* set, each set being a disjunction of mutually exclusive values.
- Disjunctive sets are in one-to-one correspondence with so-called "logical" belief functions, and probability measures are another special case of belief functions. A belief function can thus be seen both as a generalised probability measure and as a generalised set; it makes it possible to combine reasoning mechanisms from probability theory (conditioning, marginalisation), with set-theoretic operations (intersection, union, cylindrical extension, interval computations, etc.)

DS theory thus provides a very general framework allowing us to reason with imprecise and uncertain information. In particular, it makes it possible to represent states of knowledge close to total ignorance and, consequently, to model situations in which the available knowledge is too limited to be properly represented in the probabilistic formalism. Dempster's rule of combination [565] is an important building block of DS theory, because it provides a general mechanism for combining independent pieces of evidence. Recently, on extension of DS theory based on random fuzzy sets and a product-combination rule has been introduced [180, 182], allowing, in particular, to define practical models of belief functions in $\mathbb{R}^p$.

The first applications of DS theory to machine learning date back to the 1990's and concerned classifier combination [651, 530], each classifier being considered as a piece of evidence and combined by Dempster's rule (see, e.g., [515] for a refinement of this idea taking into account the dependence between classifier outputs). In [176], Denœux combined Shafer's idea of evidence combination with distance-based classification to introduce the evidential K-NN classifier [176]. In this method, each neighbour of an instance to be classified is considered as a piece of evidence about the class of that instance and is represented by a belief function. The K belief functions induced by the K nearest neighbours are then combined by Dempster's rule. Extensions of this simple scheme were later proposed in [680, 406, 186]. An neural network version based on prototypes was introduced in [177] and combined with deep architectures in [599, 598]. A neural network model for regression based on similar ideas and quantifying prediction uncertainty using random fuzzy sets was recently described in [181].

The evidential K-NN rule is, thus, the first example of an "evidential classifier". Typically, an evidential classifier breaks down the evidence of each input feature vector into elementary mass functions and combines them by Dempster's rule. The combined mass function can then be used for decision-making. Thanks to the generality and expressiveness of the belief function formalism, evidential classifiers provide more informative outputs than those of conventional classifiers. This expressiveness can be exploited, in particular, for uncertainty quantification, novelty detection and information fusion in decision-aid or fully automatic decision systems [186].

In [186], it is shown that not only distance-based classifiers such as the evidential K-NN rule, but also a broad class of supervised machine learning algorithms, can be seen as evidential classifiers. This class contains logistic regression and its non linear generalizations, including multilayer feedforward neural networks, generalized additive models, support vector machines and, more generally, all classifiers based on linear combinations of inputs or higher-order features and their transformation through the logistic or softmax transfer function. Such *generalized logistic regression classifiers* can be seen as combining elementary pieces of evidence supporting each class or its complement using Dempster's rule. The output class probabilities are then normalized plausibilities according to some underlying belief function. This "hidden" belief function provides a more informative description of the classifier output than the class probabilities, and can be used for decision-making. Also, the individual belief functions computed by each of the features provide insight into the internal operation of classifier and can help interpret its decisions. This finding opens a new perspective for the study and practical applications of a wide range of machine learning algorithms.

9.3 Maximum Likelihood Under Coarse Data

When data is missing or just imprecise (one then speaks of *coarse data*), statistical methods need to be adapted. In particular, the question is whether one wishes to model the observed phenomenon *along* *with* the limited precision of observations or *despite* such imprecision. The latter view comes down to completing the data in some way (using imputation methods). A well-known method that does it is the EM algorithm [175]. This technique makes strong assumptions on the measurement process so as to relate the distribution ruling the underlying phenomenon and the one ruling the imprecise outcomes. The EM algorithm and its extensions is extensively used for clustering (using Gaussian mixtures) and learning Bayesian nets.

However, the obtained result, where by virtue of the algorithm, data has become complete and precise, is not easy to interpret. If we want to be faithful to the data and its imperfections, one way is to build a model that accounts for the imprecision of observations, i.e., a set-valued model. This is the case if a belief function is obtained via maximum likelihood on imprecise observations: one optimises the so-called *visible likelihood function* [136]. The idea is to cover all precise models that could have been derived, had the data been precise. Imprecise models are useful to lay bare ignorance when it is present, so as to urge finding more data, but it may be problematic for decision or prediction problems, when we have to act or select a value despite ignorance.

Ideally we should optimize the likelihood function based on the actual values hidden behind the imprecise observations. But such a likelihood function is ill-known in the case of coarse data [136]. In that case, we can adopt several strategies:

- We can make assumptions on the measurement process so as to create a tight link between the hidden likelihood function pertaining to the outcomes of the real phenomenon, and the visible likelihood of the imprecise observations, for instance the CAR (coarsening at random) assumption [305], or the superset assumption [331]. In that case, the coarseness of the data can be in some sense ignored. See [349] for a general discussion.

- Another approach is to pick a suitable hidden likelihood function among the ones compatible with the imprecise data, for instance using an optimistic maximax approach that considers that the true sample is the best possible sample in terms of likelihood compatible with the imprecise observation [330]. This approach chooses a compatible probability distribution with minimal entropy, hence tends to disambiguate the data. On the contrary, the maximin approach considers that the true sample is the worst compatible sample in terms of likelihood. This approach chooses a compatible probability distribution with maximal entropy. Those two approaches adopt extreme points of view on the entropy of the probability distribution. More recently, an approach based on the likelihood ratio that maximizes the minimal possible likelihood ratio over the compatible probability distributions is proposed in [290]. This method achieves a trade-off between the previous two more extreme approaches and is able to quantify the quality of the chosen probability distribution in regards to all possible ones. In these approaches, the measurement process is ignored.

- Yet another approach is to extend the notion of likelihood to the case of imprecise and uncertain data. In the classical case of precise data, the likelihood can be seen as a measure of agreement between the data and a model. In [178], Denœux generalized this idea to the case of imprecise and uncertain data by defining the likelihood as one minus the degree of conflict (in the sense of Dempster-Shafer theory) between the probabilistic model and the data, represented by belief functions. An extension of the EM algorithm, called the Evidential EM (E^2M) algorithm, allows us to maximize this generalized likelihood. This approach has been applied, e.g., to classification with soft labels [514], partially supervised independent factor analysis [117] and hidden Markov models with partial knowledge of hidden states [520].

See [137, 332, 192] for more discussions about such methods for statistical inference with poor quality data.

Besides, another line of work for taking into account the scarcity of data in ML is to use a new cumulative entropy-like function that together considers the entropy of the probability distribution and the uncertainty pertaining to the estimation of its parameters. It takes advantage of the ability of a possibility distribution to upper bound a family of probabilities previously estimated from a limited set of examples [562, 563], and of the link between possibilistic specificity order (fuzzy set inclusion) and probabilistic entropy [205]. This approach enables the expansion of decision trees to be limited when the number of examples at the current final nodes is too small. Similar ideas can also be found in the imprecise probability literature, where upper entropies over sets of probabilities are used to limit the growth of the tree [3], and it would not be surprising to find strong formal links between the two approaches.

9.4 The EM Algorithm and Belief Revision

Injecting concepts from KRR, when explaining the EM algorithm, may help better figure out what it does. In the most usual case, coarse data are elements of a partition of the domain of values of some hidden variable. Given a class of parametric statistical models, the idea is to iteratively construct a precise model that fits the data as much as possible, by first generating at each step a precise observation sample in agreement with the incomplete data, followed by the computation of a new model obtained by applying the maximum likelihood method to the last precise sample. These two steps are repeated until convergence to a model is achieved.

In [135], it has been shown that the observation sample implicitly built at each step can be represented by a probability distribution on the domain of the hidden variable that is in agreement with the observed frequencies of the coarse data. It is obtained by applying, at each step of the procedure, the oldest (probabilistic) revision rule well-known in AI and epistemology, namely Jeffrey's rule [355], to the current best parametric model. This form of revision considers a prior probability $p(x,\theta)$ on the domain $\mathscr{X}$ of a variable X, and new information made of a probability distribution over a partition $\{A_1,A_2,\dots,A_n\}$ of $\mathscr{X}$ (representing the coarse data). If p'_i is the "new" probability of A_i, the old distribution $p(x,\theta)$ is revised so as to be in agreement with the new information. The revised probability function is of the form $p'(x,\theta) = \sum_{i=1}^{n} p'_i \cdot p(x,\theta|A_i)$. The revision step minimally changes the prior probability function in the sense of Kullback-Leibler relative entropy.

In the case of the EM algorithm, p'_i is the frequency of the coarse observation A_i, and $p(x,\theta)$ is the current best parametric model. The distribution $p'(x,\theta)$ corresponds to a new sample of X in agreement with the coarse observation. In other words, the EM algorithm revises the parametric model so as to make it consistent with the coarse data, minimizing the relative (entropic) distance between the current parametric model and the new probability distribution in agreement with the coarse data, then applies maximum likelihood to the new obtained sample, so as to get a new parametric model, and so on, till convergence is attained.

Note that other recent works have also bridged belief change with classification [130, 553].

9.5 Possibility theory in Statistics

Possibility theory [208] is the simplest non-additive uncertainty model devoted to incomplete information. It has several formal frameworks, ranging from purely ordinal representations to numerical settings. Numerical possibility theory has connections to statistics [204], hence to machine learning.

Possibility measures ideally account for imprecise but coherent data (e.g. a bunch of nested intervals), while probability theory is tailored to precise randomly scattered data. In general, data are neither precise nor nested, and possibility theory can be used as an approximate representation of such imprecision-tainted data. Interestingly possibilistic representations also fit the situations when precise data is scarce, often obtaining non-parametric models using probabilistic inequalities. See [212] for various techniques for representing data using possibility distributions.

Possibility distributions are also a powerful tool to represent the dispersion of probability distributions [448]. This is done by transforming probability distributions into possibility distributions and pointwisely comparing the latter. This way of comparing the peakedness of probability distributions has been proved to have some connections with entropy[205].

Finally, possibility theory can be used in inferential statistics, noticing that a likelihood function is a kind of possibility distribution. In particular, the maximum likelihood principle can be seen as an application of possibility theory. Besides there is a counterpart of Bayes theorem in possibility theory. As a consequence, we can adapt the maximum likelihood approach to conditional distributions as well. While such formal links have been known since a long time [623], Denœux [180, 182] has recently taken this idea a step further and has shown that statistical predictions can be made by combining knowledge about the parameters given by the relative likelihood function seen as a possibility distribution, with a probability distribution representing randomness. The result is a random fuzzy set that describes prediction uncertainty. This approach has been applied to logistic regression in [183].

For more details, the reader is referred to the survey paper by Dubois and Prade on possibility theory and learning in this special issue [215].

9.6 Beyond possibility theory: learning an imprecise linear model.

The simplest regression models associating input/output pairs are linear. A linear model can be formulated in probabilistic terms, where the output is an expectation of the inputs with respect to a probability measure. In such a case, estimating the weight associated with each input is enough to determine the model. To reflect the difficulty of estimating these weights from a finite training set, or to impact the potential imprecision of the observations, it may be interesting to replace the probabilistic model with a Choquet capacity defining a credal set, i.e. the set of probability measures that are dominated by the Choquet capacity. However, as mentioned in Section 9.1.1, such a model can be computationally very challenging. Modeling the credal set by a possibility measure is a good compromise [204] since it suffices, as in the probabilistic case, to estimate the weight associated with each input. One of the particularities of such a model is that its output is imprecise, as being the set of outputs obtained by an expectation with respect to any probability belonging to the credal set [425]. The result is a kind of imprecise linear model. However, this approach only concerns processes whose precise model is a probabilistic expectation.

In [587], it is proposed to extend this approach to any process that can be modeled by a linear combination. It uses a non-monotonic extension of the possibility measures, the MacSum set function, where the weights associated to each input can be positive or negative. The MacSum set function is the sum of a maxitive function associated with positive weights and a minitive function associated with negative weights. This structure makes it possible to define a convex set of linear combinations whose sum of weights is identical, i.e. an imprecise linear model. Learning such a model can be carried out in regression form [311], as in the precise linear case, and admits both precise and imprecise data [586]. In the case where the inputs are imprecise, two extreme situations can arise: either the imprecision of the inputs and the model add up (disjunctive modeling) or they compensate for each other (conjunctive modeling).

10 Conclusion

KRR and ML, the two main areas of AI, have been developed rather independently for several decades. As a consequence, most of the researchers in one area are, to a large extent, ignorant of what has been going on in the other area. The intended purpose of this joint work is to provide an inventory of the meeting points between KRR and ML lines of research. The paper has reviewed some concerns that are shared by the two areas, maybe in different ways, surveyed various paradigms that are at the border between KRR and ML and provided an overview of different hybridizations of KRR and ML tools. The works covered may be old or recent, well-known as well as overlooked.

By the breadth of subjects covered, this article is a real challenge. However, even if the paper has been substantially revised, refreshed and expanded since its original version was made available [86], some subsections may still be found too sketchy. There is absolutely no claim of completeness of any kind, not even of being fully up to date. Here are just a few examples of topics not addressed or poorly handled in the article: namely, reinforcement learning, ontology representation and learning [27] (and more generally the learning of graphical representations), argumentation and ML [13], or the logical analysis of data [80] (see also [118, 459]). Moreover, each topic covered in this paper is only outlined and would deserve to be discussed in further details. Even if the current list of references is lengthy, it is certainly incomplete and not ideally balanced. However, this work is complemented by the articles in the same volume, which are briefly introduced in this article and which develop some of the issues raised, or address topics not covered, while providing additional references.

The aim of this paper is to help facilitating the understanding between researchers in the two areas, with a perspective of cross-fertilisation and mutual benefits. Yet, we should be aware that the mathematics of ML and the mathematics of KRR are quite different if we consider the main trends in each area. In ML the basic paradigm is a matter of approximating functions in continuous spaces (which then calls for optimization in continuous spaces). The mathematics of ML are close to those of signal processing and automatic control (as pointed out in [395]]), while KRR is dominated by logic and discrete mathematics, leading to an – at least apparent – opposition between geometry and logic [431][9]. But functions also underlie KRR, once one notices that a set of (fuzzy or weighted) rules is like an aggregation function [210], whose computation may take the form of a generalized matrix calculus ('generalized' in the sense that the operations are not necessarily restricted to sum and product) [36]. Let us also note that the convolution of functions (a key tool in signal processing) is no longer restricted to a linear, sum/product-based setting [466]. Besides, let us note that deep learning methods are now able to predict Boolean functions, e.g., [151], which is another clue that the gap between Logic and ML is diminishing.

ML methods are to a large extent highly quantitative (artificial neurons remain quantitative), while KRR approaches are often qualitative. Even if being quantitative often contributes to good performance, it may be of interest of also developing more qualitative views to facilitate the interface with KRR methods, just as "System 1" and "System 2" (in the sense of [358]) interact in human beings.

Examples of states of fact that might call for some cooperation between ML and KRR are for instance, as low shot learning, unsupervised learning of disentangled representations is fundamentally impossible without inductive biases on both the models and the data [422]; the local explanation methods for deep neural networks lack sensitivity to parameter values [6]; when trained on one task, then trained on a second task, many machine learning models "forget" how to perform the first task [271].

Generally speaking, dealing with KRR projects without any ML concern may narrow the horizon of the research problems, and it seems also difficult to envisage ML without KRR in some situations. In any case, it can be observed that there are more and more research works mixing KRR and ML tools. Trustworthy AI and accountability, frugal AI [53] are important and popular issues that may benefit in the long range of better synergies between KRR and ML.

11 Acknowledgements

The authors thank the CNRS GDR "Formal and Algorithmic Aspects of Artificial Intelligence" for its support, which in 2023 became the CNRS GDR RADIA for "Raisonnement, Apprentissage, et Décision en Intelligence Artificielle" (Reasoning, Learning and Decision in AI).

References

[1] AAAI. *The 33rd AAAI Conf. on Artificial Intelligence (AAAI'19), The 31st Innovative Applications of Artificial Intelligence Conf., IAAI 2019, The 9th AAAI Symp. on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Jan. 27 - Feb. 1*. AAAI Press, 2019.

[2] A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Commun.*, 7(1):39–59, 1994.

[3] J. Abellán and S. Moral. Building classification trees using the total uncertainty criterion. *Int. J. of Intelligent Systems*, 18(12):1215–1225, 2003.

[4] A. Abramé and D. Habet. AHMAXSAT : Description and evaluation of a branch and bound Max-SAT solver. *Journal on Satisfiability, Boolean Modeling and Computation*, 2015.

[5] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[6] J. Adebayo, J. Gilmer, I. J. Goodfellow, and B. Kim. Local explanation methods for deep neural networks lack sensitivity to parameter values. In *Workshop Track Proc. 6th Int. Conf. on Learning Representations, ICLR'18, Vancouver, April 30 - May 3*. OpenReview.net, 2018.

[9] See also the conference https://www.youtube.com/watch?v=BpX890StRvs .

[7] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In *Proc. 32nd Int. Conf. on Neural Information Processing Systems*, NIPS'18, page 9525–9536, Red Hook, NY, USA, 2018. Curran Associates Inc.

[8] G. Aglin, S. Nijssen, and P. Schaus. Learning optimal decision trees under memory constraints. In M. Amini, S. Canu, A. Fischer, T. Guns, P. K. Novak, and G. Tsoumakas, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2022, Grenoble, France, September 19-23, 2022, Proceedings, Part V*, volume 13717 of *LNCS*, pages 393–409. Springer, 2022.

[9] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. SIGMOD Conf, Washington, DC, May 26-28*, pages 207–216. ACM Press, 1993.

[10] Y. C. C. Alarcón and S. Destercke. Distributionally robust, skeptical binary inferences in multi-label problems. In *Int. Symp. on Imprecise Probability: Theories and Applications*, pages 51–60. PMLR, 2021.

[11] J. M. Alonso, C. Castiello, and C. Mencar. A bibliometric analysis of the explainable artificial intelligence research field. In J. Medina, M. Ojeda-Aciego, J. L. V. Galdeano, D. A. Pelta, I. P. Cabrera, B. Bouchon-Meunier, and R. R. Yager, editors, *Proc. 17th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'18, Cádiz, June 11-15, Part I*, volume 853 of *Communications in Computer and Information Science*, pages 3–15. Springer, 2018.

[12] L. Amgoud and J. Ben-Naim. Axiomatic foundations of explainability. In *Proc. of the 23st Int. Conf. on Artificial Intelligence, IJCAI22*, pages 636–642, 2022.

[13] L. Amgoud and M. Serrurier. Agents that argue and explain classifications. *Autonomous Agents and Multi-Agent Systems*, 16(2):187–209, 2008.

[14] S. Amizadeh, S. Matusevych, and M. Weimer. Learning to solve Circuit-SAT: An unsupervised differentiable approach. In *Proc. 7th Int. Conf. on Learning Representations (ICLR'19), New Orleans, May 6-9*. OpenReview.net, 2019.

[15] S. Amizadeh, S. Matusevych, and M. Weimer. PDP: A general neural framework for learning constraint satisfaction solvers. *CoRR*, abs/1903.01969, 2019.

[16] M. Amoussou, K. Belahcène, C. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane. Questionable stepwise explanations for a robust additive preference model. *Int. J. of Approximate Reasoning*, this issue:108982, 2023.

[17] M. Ancona, E. Ceolini, A. C. Öztireli, and M. H. Gross. A unified view of gradient-based attribution methods for deep neural networks. *CoRR*, 2017.

[18] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin. Learning certifiably optimal rule lists. In *Proc. 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, Halifax, Aug. 13 - 17*, pages 35–44. ACM, 2017.

[19] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin. Learning certifiably optimal rule lists for categorical data. *J. Mach. Learn. Res.*, 18:234:1–234:78, 2017.

[20] A. N. Angelopoulos, S. Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

[21] D. Angluin. Queries and concept learning. *Mach. Learn.*, 2(4):319–342, 1987.

[22] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling. Explainable agents and robots: Results from a systematic literature review. In *AAMAS*, pages 1078–1088, 2019.

[23] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 2009.

[24] A. Antonucci and G. Corani. Likelihood-based naive credal classifier. In *Int. Symp. on Imprecise Probability: Theories and Applications (ISIPTA)*, volume 11, pages 21–30. Citeseer, 2011.

[25] A. Antonucci and G. Corani. The multilabel naive credal classifier. *Int. J. of Approximate Reasoning*, 83:320–336, 2017.

[26] R. Arcangioli, C. Bessiere, and N. Lazaar. Multiple constraint acquisition. In S. Kambhampati, editor, *Proc. 25th Int. Joint Conf. on Artificial Intelligence (IJCAI'16), New York, July 9-15*, pages 698–704. IJCAI/AAAI Press, 2016.

[27] R. Arp, B. Smith, and A. Spear. *Building Ontologies with Basic Formal Ontology*. MIT Press, 2015.

[28] Z. Assaghir, M. Kaytoue, and H. Prade. A possibility theory-oriented discussion of conceptual pattern structures. In A. Deshpande and A. Hunter, editors, *Proc. 4th Int. Conf. (SUM'10)on Scalable Uncertainty Management Toulouse, Sept. 27-29*, volume 6379 of *LNCS*, pages 70–83. Springer, 2010.

[29] J. Atif, I. Bloch, and C. Hudelot. Some relationships between fuzzy sets, mathematical morphology, rough sets, f-transforms, and formal concept analysis. *Int.J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 24, No. Suppl. 2:1–32, 2016.

[30] A. M. Aung, Y. Fadila, R. Gondokaryono, and L. Gonzalez. Building robust deep neural networks for road sign detection. *CoRR*, abs/1712.09327, 2017.

[31] M. Ayel and M.-C. Rousset. *La Cohérence dans les Bases de Connaissances*. Cepadues, 1990.

[32] F. Baader, S. Brandt, and C. Lutz. Pushing the EL envelope. In L. P. Kaelbling and A. Saffiotti, editors, *Proc. 19th Int. Joint Conf. on Artificial Intelligence (IJCAI'05), Edinburgh, July 30 - Aug. 5*, pages 364–369, 2005.

[33] M. Baader, M. Mirman, and M. T. Vechev. Universal approximation with certified networks. *CoRR*, abs/1909.13846, 2019.

[34] I. Baaj. *Explainability of possibilistic and fuzzy rule-based systems*. PhD thesis, Sorbonne Université, 2022.

[35] I. Baaj. Learning rule parameters of possibilistic rule-based system. In *Proc. 2022 IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE, 2022.

[36] I. Baaj. On the handling of inconsistent systems of max-min fuzzy relational equations. *Fuzzy Sets and Systems*, page 108912, 2024.

[37] I. Baaj, J. Poli, W. Ouerdane, and N. Maudet. Min-max inference for possibilistic rule-based system. In *Proc. 30th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE'21), Luxembourg, July 11-14*, pages 1–6. IEEE, 2021.

[38] I. Baaj, J.-P. Poli, W. Ouerdane, and N. Maudet. Representation of explanations of possibilistic inference decisions. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 16th European conf., ECSQARU 2021, Prague, Czech Republic, September 21–24, 2021, Proceedings 16*, pages 513–527. Springer, 2021.

[39] B. Babaki, T. Guns, and S. Nijssen. Constrained clustering using column generation. In *Integration of AI and OR Techniques in Constraint Programming - Proc. 11th Int. Conf. CPAIOR'14, Cork, Ireland, May 19-23*, pages 438–454, 2014.

[40] S. H. Bach, M. Broecheler, B. Huang, and L. Getoor. Hinge-loss Markov random fields and Probabilistic Soft Logic. *J. of Machine Learning Research*, 18:109:1–109:67, 2017.

[41] J. Baget, S. Benferhat, Z. Bouraoui, M. Croitoru, M. Mugnier, O. Papini, S. Rocher, and K. Tabia. Inconsistency-tolerant query answering: Rationality properties and computational complexity analysis. In L. Michael and A. C. Kakas, editors, *Proc. 15th Europ. conf. on Logics in Artificial Intelligence (JELIA'16), Larnaca, Cyprus, Nov. 9-11*, volume 10021 of *LNCS*, pages 64–80, 2016.

[42] M. Balabanović and Y. Shoham. Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.

[43] C. Balkenius and P. Gärdenfors. Nonmonotonic inferences in neural networks. In *Proc. 2nd Int. Conf. on Princip. of Knowl. Represent. and Reas. (KR'91). Cambridge, MA*, pages 32–39, 1991.

[44] M. Balunovic, P. Bielik, and M. T. Vechev. Learning to solve SMT formulas. In Bengio et al. [60], pages 10338–10349.

[45] T. Baluta, S. Shen, S. Shinde, K. S. Meel, and P. Saxena. Quantitative verification of neural networks and its security applications. In L. Cavallaro, J. Kinder, X. Wang, and J. Katz, editors, *Proc. of the 2019 ACM SIGSAC Conf. on Computer and Communications Security (CCS'19), London, Nov. 11-15*, pages 1249–1264. ACM, 2019.

[46] K. Bansal, S. M. Loos, M. N. Rabe, C. Szegedy, and S. Wilcox. HOList: An environment for machine learning of higher order logic theorem proving. In *Proc 36th Int. Conf. on Machine Learning (ICML'19), Long Beach June 9-15*, pages 454–463, 2019.

[47] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82,Äì115, 2020.

[48] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. In J. W. Lloyd, V. Dahl, U. Furbach, M. Kerber, K. Lau, C. Palamidessi, L. M. Pereira, Y. Sagiv, and P. J. Stuckey, editors, *Computational Logic - CL 2000, Proc. 1st Int. Conf., London, 24-28 July*, volume 1861 of *LNCS*, pages 972–986. Springer, 2000.

[49] K. Belahcene, C. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane. Explaining robust additive utility models by sequences of preference swaps. *Theory and Decision*, 82(2):151–183, 2017.

[50] N. Beldiceanu and H. Simonis. A model seeker: Extracting global constraint models from positive examples. In M. Milano, editor, *Proc. 18th Int. Conf. on Principles and Practice of Constraint Programming (CP'12), Québec City, Oct. 8-12*, volume 7514 of *LNCS*, pages 141–157. Springer, 2012.

[51] I. Bello, H. Pham, Q. V. Le, M. Norouzi, and S. Bengio. Neural combinatorial optimization with reinforcement learning. In *Proc. 5th Int. Conf. on Learning Representations (ICLR'17), Toulon, Apr. 24-26*. OpenReview.net, 2017.

[52] R. Belohlavek. *Fuzzy Relational Systems. Foundations and Principles*. Kluwer, 2002.

[53] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In M. C. Elish, W. Isaac, and R. S. Zemel, editors, *FAccT '21: 2021 ACM conf. on Fairness, Accountability, and Transparency, Virtual Event / Toronto, March 3-10, 2021*, pages 610–623. ACM, 2021.

[54] S. Benferhat, D. Dubois, L. Garcia, and H. Prade. On the transformation between possibilistic logic bases and possibilistic causal networks. *Int. J. Approx. Reasoning*, 29(2):135–173, 2002.

[55] S. Benferhat, D. Dubois, S. Lagrue, and H. Prade. A big-stepped probability approach for discovering default rules. *Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(Supplement-1):1–14, 2003.

[56] S. Benferhat, D. Dubois, and H. Prade. Some syntactic approaches to the handling of inconsistent knowledge bases: A comparative study. Part 1: The flat case. *Studia Logica*, 58(1):17–45, 1997.

[57] S. Benferhat, D. Dubois, and H. Prade. An overview of inconsistency-tolerant inferences in prioritized knowledge bases. In D. Dubois, H. Prade, and E. P. Klement, editors, *Fuzzy Sets, Logics and Reasoning about Knowledge*, pages 395–417. Springer Netherlands, Dordrecht, 1999.

[58] S. Benferhat, D. Dubois, and H. Prade. The possibilistic handling of irrelevance in exception-tolerant reasoning. *Ann. Math. Artif. Intell.*, 35(1-4):29–61, 2002.

[59] S. Benferhat and K. Tabia. Inference in possibilistic network classifiers under uncertain observations. *Ann. Math. Artif. Intell.*, 64(2-3):269–309, 2012.

[60] S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors. *Advances in Neural Information Processing Systems 31: Annual Conf. on Neural Information Processing Systems (NeurIPS'18), Dec. 3-8, Montréal*, 2018.

[61] Y. Bengio, A. Lodi, and A. Prouvost. Machine learning for combinatorial optimization: a methodological tour d'horizon. *CoRR*, abs/1811.06128, 2018.

[62] Y. Bengio and N. Malkin. Machine learning and information theory concepts towards an ai mathematician. *arXiv preprint arXiv:2403.04571*, 2024.

[63] P. Besnard and A. Hunter. *Reasoning with Actual and Potential Contradictions*. Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol.2, (D. Gabbay and Ph. Smets, series editors. Kluwer Acad. Publ., 1998.

[64] T. R. Besold, A. S. d'Avila Garcez, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K. Kühnberger, L. C. Lamb, P. M. V. Lima, L. de Penning, G. Pinkas, H. Poon, and G. Zaverucha. Neural-symbolic learning and reasoning: A survey and interpretation. In P. Hitzler and M. K. Sarker, editors, *Neuro-Symbolic Artificial Intelligence: The State of the Art*, volume 342 of *Frontiers in Artificial Intelligence and Applications*, pages 1–51. IOS Press, 2021. Also in CoRR/abs/1711.03902 (2017).

[65] C. Bessiere, C. Carbonnel, A. Dries, E. Hebrard, G. Katsirelos, N. Narodytska, C. Quimper, K. Stergiou, D. C. Tsouros, and T. Walsh. Learning constraints through partial queries. *Artif. Intell.*, 319:103896, 2023.

[66] C. Bessiere, C. Carbonnel, and A. Himeur. Learning constraint networks over unknown constraint languages. In *Proc. 32nd Int. Joint Conf. on Artificial Intelligence (IJCAI'23), Macao, Aug. 19th-25th*, pages 1876–1883. ijcai.org, 2023.

[67] C. Bessiere, R. Coletta, E. Hebrard, G. Katsirelos, N. Lazaar, N. Narodytska, C. Quimper, and T. Walsh. Constraint acquisition via partial queries. In F. Rossi, editor, *Proc. 23rd Int. Joint Conf. on Artificial Intelligence (IJCAI'13), Beijing, Aug. 3-9*, pages 475–481. IJCAI/AAAI, 2013.

[68] C. Bessiere, R. Coletta, F. Koriche, and B. O'Sullivan. A sat-based version space algorithm for acquiring constraint satisfaction problems. In J. Gama, R. Camacho, P. Brazdil, A. Jorge, and L. Torgo, editors, *Proc. 16th Europ. Conf. on Machine Learning (ECML'05), Porto, Oct. 3-7*, volume 3720 of *LNCS*, pages 23–34. Springer, 2005.

[69] C. Bessiere, R. Coletta, B. O'Sullivan, and M. Paulin. Query-driven constraint acquisition. In M. M. Veloso, editor, *Proc. 20th Int. Joint Conf. on Artificial Intelligence (IJCAI'07), Hyderabad, Jan.6-12*, pages 50–55, 2007.

[70] C. Bessiere, E. Hebrard, and B. O'Sullivan. Minimising decision tree size as combinatorial optimisation. In I. P. Gent, editor, *Principles and Practice of Constraint Programming - CP 2009, Proc. 15th Int. Conf. CP'09, Lisbon, Sept. 20-24*, volume 5732 of *LNCS*, pages 173–187. Springer, 2009.

[71] C. Bessiere, F. Koriche, N. Lazaar, and B. O'Sullivan. Constraint acquisition. *Artif. Intell.*, 244:315–342, 2017.

[72] S. Bhatia, P. Kohli, and R. Singh. Neuro-symbolic program corrector for introductory programming assignments. In M. Chaudron, I. Crnkovic, M. Chechik, and M. Harman, editors, *Proc. 40th Int. Conf. on Software Engineering (ICSE'18), Gothenburg, May 27 - June 3*, pages 60–70. ACM, 2018.

[73] O. Biran and C. Cotton. Explanation and justification in machine learning: A survey. In *Proc. of the Workshop on eXplainable Artificial Intelligence, XAI@IJCAI*, volume 8, page 1, 2017.

[74] J. Błaszczyński, R. Słowiński, and M. Szelag. Sequential covering rule induction algorithm for variable consistency rough set approaches. *Information Sciences*, 181(5):987 – 1002, 2011.

[75] S. Blum, R. Koudijs, A. Ozaki, and S. Touileb. Learning horn envelopes via queries from language models. *Int. J. of Approximate Reasoning*, this issue:109026, 2023.

[76] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37:1719‚Äì1778, 2023.

[77] R. Bommasani, D. Hudson, E. Adeli, R. Altman, S. Arora, S. Arx, M. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Davis, D. Demszky, and P. Liang. On the opportunities and risks of foundation models, 08 2021.

[78] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013.

[79] S.-E. Bornscheuer, S. Hölldobler, Y. Kalinke, and A. Strohmaier. Massively parallel reasoning. In W. Bibel and P. H. Schmitt, editors, *Automated Deduction - A Basis for Applications. Volume II: Systems and Implementation Techniques*, pages 291–321. Springer, 1998.

[80] E. Boros, Y. Crama, P. L. Hammer, T. Ibaraki, A. Kogan, and K. Makino. Logical analysis of data: classification with justification. *Annals OR*, 188(1):33–61, 2011.

[81] M. Bounhas, M. Pirlot, and H. Prade. Predicting preferences by means of analogical proportions. In M. Cox, P. Funk, and S. Begum, editors, *Proc. 26th Int. Conf. on Case-Based Reason- ing Research (ICCBR'18), Stockholm, July 9-12*, volume 11156 of *LNCS*, pages 515–531. Springer, 2018.

[82] M. Bounhas, M. Pirlot, H. Prade, and O. Sobrie. Comparison of analogy-based methods for predicting preferences. In N. B. Amor, B. Quost, and M. Theobald, editors, *Proc. 13th Int. Conf on Scalable Uncertainty Management (SUM'19), Compiègne, France, December 16-18*, volume 11940 of *LNCS*, pages 339–354. Springer, 2019.

[83] M. Bounhas and H. Prade. Analogy-based classifiers: An improved algorithm exploiting competent data pairs. *Int. J. Approx. Reason.*, 158:108923, 2023.

[84] M. Bounhas and H. Prade. Revisiting analogical proportions and analogical inference. *Int. J. of Approximate Reasoning*, this issue:????, 2024.

[85] M. Bounhas, H. Prade, and G. Richard. Analogy-based classifiers for nominal or numerical data. *Int. J. Approx. Reasoning*, 91:36–55, 2017.

[86] Z. Bouraoui, A. Cornuéjols, T. Denoeux, S. Destercke, D. Dubois, R. Guillaume, J. Marques-Silva, J. Mengin, H. Prade, S. Schockaert, M. Serrurier, and C. Vrain. From shallow to deep interactions between knowledge representation, reasoning and machine learning (Kay R. Amel group). *CoRR*, abs/1912.06612, 2019.

[87] Z. Bouraoui and S. Schockaert. Learning conceptual space representations of interrelated concepts. In *Proc. 27th Int. Joint Conf. on Artificial Intelligence*, pages 1760–1766, 2018.

[88] C. Boutilier, R. I. Brafman, C. Domshlak, H. H. Hoos, and D. Poole. CP-nets: a tool for representing and reasoning with conditional ceteris paribus preference statements. *J. of Artificial Intelligence Research*, 21:135–191, 2004.

[89] Q. Brabant, M. Couceiro, D. Dubois, H. Prade, and A. Rico. Extracting decision rules from qualitative data via Sugeno utility functionals. In J. Medina, M. Ojeda-Aciego, J. L. V. Galdeano, D. A. Pelta, I. P. Cabrera, B. Bouchon-Meunier, and R. R. Yager, editors, *Proc. 17th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'18), Cádiz, June 11-15, Part I*, volume 853 of *CCIS*, pages 253–265. Springer, 2018.

[90] G. Brewka, V. W. Marek, and M. Truszczynski, editors. *Nonmonotonic Reasoning. Essays Celebrating Its 30th Anniversary*, volume 31 of *Studies in Logic*. College Publication, 2011.

[91] C. Brix, M. N. Müller, S. Bak, T. T. Johnson, and C. Liu. First three years of the international verification of neural networks competition (VNN-COMP). *Int. J. Softw. Tools Technol. Transf.*, 25(3):329–339, 2023.

[92] S. Bromberger, editor. *On What we Know we Don't Know: Explanation, Theory, Linguistics, and How Questions Shape Them*. University of Chicago Press, 1992.

[93] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[94] R. Bunel, M. J. Hausknecht, J. Devlin, R. Singh, and P. Kohli. Leveraging grammar and reinforcement learning for neural program synthesis. In *Proc. 6th Int. Conf. on Learning Representations (ICLR'18), Vancouver, Apr. 30 - May 3*. OpenReview.net, 2018.

[95] R. Bunel, J. Lu, I. Turkaslan, P. H. S. Torr, P. Kohli, and M. P. Kumar. Branch and bound for piecewise linear neural network verification. *CoRR*, abs/1909.06588, 2019.

[96] R. Bunel, I. Turkaslan, P. H. S. Torr, P. Kohli, and M. P. Kumar. Piecewise linear neural network verification: A comparative study. *CoRR*, abs/1711.00455, 2017.

[97] R. Bunel, I. Turkaslan, P. H. S. Torr, P. Kohli, and P. K. Mudigonda. A unified view of piecewise linear neural network verification. In *Advances in Neural Information Processing Systems 31: Annual Conf. on Neural Information Processing Systems (NeurIPS'18), 3-8 Dec., Montréal*, pages 4795–4804, 2018.

[98] N. Burkart and M. F. Huber. A survey on the explainability of supervised machine learning. *J. of Artificial Intelligence Research*, 70:245–317, 2021.

[99] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Dl-lite: Tractable description logics for ontologies. In M. M. Veloso and S. Kambhampati, editors, *Proc. 20th National Conf. on Artificial Intelligence (AAAI'05), July 9-13, Pittsburgh*, pages 602–607. AAAI Press / The MIT Press, 2005.

[100] A. Campagner. Learning from fuzzy labels: Theoretical issues and algorithmic solutions. *Int. J. of Approximate Reasoning*, this issue:108969, 2023.

[101] A. Campagner, D. Ciucci, and T. Denœux. Belief functions and rough sets: Survey and new insights. *Int. J. of Approx. Reas.*, 143:192–215, 2022.

[102] M. Canabal-Juanatey, J. M. Alonso-Moral, A. Catala, and A. Bugarín-Diz. Enriching interactive explanations with fuzzy temporal constraint networks. *Int. J. of Approximate Reasoning*, this issue:109128, 2024.

[103] L. Cardelli, M. Kwiatkowska, L. Laurenti, N. Paoletti, A. Patane, and M. Wicker. Statistical guarantees for the robustness of bayesian neural networks. In S. Kraus, editor, *Proc. 28th Int. Joint Conf. on Artificial Intelligence (IJCAI'19), Macao, Aug. 10-16*, pages 5693–5700. ijcai.org, 2019.

[104] W. Carnielli and M. Coniglio. *Paraconsistent Logic: Consistency, Contradiction and Negation*. Springer, 2016.

[105] J. Castro. Fuzzy logic controllers are universal approximators. *IEEE Trans. Systems, Man, and Cybernetics*, 25:629 – 635, 1995.

[106] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay. Adversarial attacks and defences: A survey. *CoRR*, abs/1810.00069, 2018.

[107] P. Chalasani, S. Jha, A. Sadagopan, and X. Wu. Adversarial learning and explainability in structured datasets. *CoRR*, abs/1810.06583, 2018.

[108] M.-W. Chang, L.-A. Ratinov, D. Roth, and V. Srikumar. Importance of semantic representation: Dataless classification. In *Proc.of the 23rd AAAI Conf.on Artificial Intelligence (AAAI'08), Chicago, July 13-17*, volume 2, pages 830–835, 2008.

[109] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, 2006.

[110] L. Charnay, J. Dibie, and S. Loiseau. Validation and explanation. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research. Vol. 1: Knowledge Representation, Reasoning and Learning*. Springer, 2019.

[111] M. Chein and M.-L. Mugnier. *Graph-based Knowledge Representation: Computational Foundations of Conceptual Graphs*. Springer, Princeton, N.J., 2009.

[112] D. Chen, Y. Bai, W. Zhao, S. Ament, J. M. Gregoire, and C. P. Gomes. Deep reasoning networks: Thinking fast and slow. *CoRR*, abs/1906.00855, 2019.

[113] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

[114] Y. Chen, H. Huang, and A. Darwiche. Towards an effective practice of learning from data and knowledge. *Int. J. of Approximate Reasoning*, this issue:109188, 2024.

[115] Z. Chen and Z. Yang. Graph neural reasoning may fail in certifying boolean unsatisfiability. *CoRR*, abs/1909.11588, 2019.

[116] W. Cheng, E. Hüllermeier, W. Waegeman, and V. Welker. Label ranking with partial abstention based on thresholded probabilistic models. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conf. on Neural Information Processing Systems 2012. Proc. of a meeting held December 3-6, Lake Tahoe, Nevada*, pages 2510–2518, 2012.

[117] Z. L. Cherfi, L. Oukhellou, E. Côme, T. Denœux, and P. Aknin. Partially supervised independent factor analysis using soft labels elicited from multiple experts: Application to railway track circuit diagnosis. *Soft Computing*, 16(5):741–754, 2012.

[118] I. Chikalov, V. V. Lozin, I. Lozina, M. Moshkov, H. S. Nguyen, A. Skowron, and B. Zielosko. *Three Approaches to Data Analysis - Test Theory, Rough Sets and Logical Analysis of Data*, volume 41 of *Intelligent Systems Reference Library*. Springer, 2013.

[119] M. Chromik and A. Butz. Human-XAI interaction: A review and design principles for explanation user interfaces. In *Proc. of the 8th IFIP TC 13 Int. Conf. on Human-Computer Interaction, INTERACT21*, pages 619–640, 2021.

[120] K. Chvalovský, J. Jakubuv, M. Suda, and J. Urban. ENIGMA-NG: efficient neural and gradient-boosted inference guidance for E. In P. Fontaine, editor, *Proc. 27th Int. Conf. on Automated Deduction (CADE'19), Natal, Brazil, Aug. 27-30*, volume 11716 of *LNCS*, pages 197–215. Springer, 2019.

[121] J. Cid-Sueiro. Proper losses for learning from partial labels. In *Advances in neural information processing systems*, pages 1565–1573, 2012.

[122] R. Cohen, M. Geva, J. Berant, and A. Globerson. Crawling the internal knowledge-base of language models. In A. Vlachos and I. Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1856–1869, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[123] W. W. Cohen, F. Yang, and K. Mazaitis. TensorLog: A probabilistic database implemented using deep-learning infrastructure. *J. Artif. Intell. Res.*, 67:285–325, 2020.

[124] R. Coletta, C. Bessière, B. O'Sullivan, E. C. Freuder, S. O'Connell, and J. Quinqueton. Semi-automatic modeling by constraint acquisition. In F. Rossi, editor, *Proc. 9th Int. Conf. on Principles and Practice of Constraint Programming (CP'03), Kinsale, Ireland, Sept. 29 - Oct. 3*, volume 2833 of *LNCS*, pages 812–816. Springer, 2003.

[125] G. Corani, A. Antonucci, and M. Zaffalon. Bayesian networks with imprecise probabilities: Theory and application to classification. In *Data Mining: Foundations and Intelligent Paradigms*, pages 49–93. Springer, 2012.

[126] A. Cornuéjols. Some thoughts about transfer learning. what role for the source domain? *Int. J. of Approx. Reas.*, page 109107, 2023.

[127] A. Cornuejols, F. Koriche, and R. Nock. Statistical computational learning. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research. Vol. 1 Knowledge Representation, Reasoning and Learning*, pages 341–388. Springer-Verlag, 2020.

[128] A. Cornuejols and C. Vrain. Designing algorithms for machine learning and data mining. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research. Vol. 2 Artificial Intelligence Algorithms*, pages 339–410. Springer-Verlag, 2020.

[129] A. Cornuéjols. Reprint of: Some thoughts about transfer learning. what role for the source domain? *Int. J. of Approximate Reasoning*, this issue:109146, 2024.

[130] S. Coste-Marquis and P. Marquis. On belief change for multi-label classifier encodings. In Z. Zhou, editor, *Proc. of the 30th Int. Joint Conf. on Artificial Intelligence (IJCAI'21), Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1829–1836. ijcai.org, 2021.

[131] M. Couceiro, N. Hug, H. Prade, and G. Richard. Analogy-preserving functions: A way to extend boolean samples. In *Proc. 26th Int. Joint Conf. on Artificial Intelligence, (IJCAI'17), Melbourne, Aug. 19-25*, pages 1575– 1581, 2017.

[132] M. Couceiro, N. Hug, H. Prade, and G. Richard. Behavior of analogical inference w.r.t. Boolean functions. In *Proc. 27th Int. Joint Conf. on Artificial Intelligence, (IJCAI'18), Stockholm, July. 13-19*, pages 2057–2063, 2018.

[133] T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *J. of Machine Learning Research*, 12(May):1501–1536, 2011.

[134] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2016.

[135] I. Couso and D. Dubois. Belief revision and the EM algorithm. In J. P. Carvalho, M. Lesot, U. Kaymak, S. M. Vieira, B. Bouchon-Meunier, and R. R. Yager, editors, *Proc. 16th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU16), Part II*, volume 611 of *CCIS*, pages 279–290. Springer, 2016.

[136] I. Couso and D. Dubois. A general framework for maximizing likelihood under incomplete data. *Int. J. of Approximate Reasoning*, 93:238–260, 2018.

[137] I. Couso, D. Dubois, and E. Hüllermeier. Maximum likelihood estimation and coarse data. In *Proc 11th Int. Conf. on Scalable Uncertainty Management (SUM'17)*, volume 10564 of *LNCS*, pages 3–16. Springer, 2017.

[138] I. Couso and L. Sánchez. Machine learning models, epistemic set-valued data and generalized loss functions: An encompassing approach. *Information Sciences*, 358:129–150, 2016.

[139] F. G. Cozman. Credal networks. *Artif. Intell.*, 120(2):199–233, 2000.

[140] F. G. Cozman. Graphical models for imprecise probabilities. *Int. J. of Approximate Reasoning*, 39:167–184, 2005.

[141] F. G. Cozman. Languages for probabilistic modeling over structured domains. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research. Vol. 2 Artificial Intelligence Algorithms*, pages 247–283. Springer-Verlag, 2020.

[142] W. Dai, Q. Xu, Y. Yu, and Z. Zhou. Tunneling neural perception and logic reasoning through abductive learning. *CoRR*, abs/1802.01173, 2018.

[143] F. d'Alché-Buc, V. Andrés, and J. Nadal. Rule extraction with fuzzy neural network. *Int. J. Neural Syst.*, 5(1):1–11, 1994.

[144] T. Dao, K. Duong, and C. Vrain. Constrained clustering by constraint programming. *Artif. Intell.*, 244:70–94, 2017.

[145] T. Dao, C. Vrain, K. Duong, and I. Davidson. A framework for actionable clustering using constraint programming. In *ECAI 2016 - 22nd European conf. on Artificial Intelligence, 29 Aug.-2 Sept. 2016, The Hague*, pages 453–461, 2016.

[146] T.-B.-H. Dao and C. Vrain. A review on declarative approaches for constrained clustering. *Int. J. of Approximate Reasoning*, this issue:109135, 2024.

[147] A. Darwiche. Logic for explainable AI. In *Proc. 38th Annual ACM/IEEE Symp.on Logic in Computer Science (LICS), Boston, June 26-29*, pages 1–11, 2023.

[148] A. Darwiche and A. Hirth. On the reasons behind decisions. In *Proc. of the European conf. on Artificial Intelligence, ECAI20*, page 712,Äì720, 2020.

[149] A. Darwiche and P. Marquis. A knowledge compilation map. *J. of Artificial Intelligence Research*, 17:229–264, 2002.

[150] A. Darwiche and P. Marquis. On quantifying literals in boolean logic and its applications to explainable AI. *J. of Artificial Intelligence Research*, 72:285–328, 2021.

[151] S. d'Ascoli, S. Bengio, J. Susskind, and E. Abbé. Boolformer: Symbolic regression of logic functions with transformers, 2023.

[152] J. Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. Reuters, 2018.

[153] I. Davidson, S. S. Ravi, and L. Shamis. A sat-based framework for efficient constrained clustering. In *Proc. SIAM Int. Conf. on Data Mining (SDM'10), April 29 - May 1, Columbus, Ohio*, pages 94–105, 2010.

[154] A. S. d'Avila Garcez, K. Broda, and D. M. Gabbay. Symbolic knowledge extraction from trained neural networks: A sound approach. *Artif. Intell.*, 125(1-2):155–207, 2001.

[155] A. S. 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. *CoRR*, abs/1905.06088, 2019.

[156] A. S. d'Avila Garcez and E. Jiménez-Ruiz, editors. *Proc.of the 16th Int. Workshop on Neural-Symbolic Learning and Reasoning as part of the 2nd Int. Joint Conf. on Learning & Reasoning (IJCLR 2022), Cumberland Lodge, Windsor Great Park, Sept. 28-30*, volume 3212 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.

[157] A. S. d'Avila Garcez and L. C. Lamb. Reasoning about time and knowledge in neural symbolic learning systems. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16 (NIPS 2003)*, pages 921–928. MIT Press, 2003.

[158] A. S. d'Avila Garcez, L. C. Lamb, and D. M. Gabbay. Neural-symbolic intuitionistic reasoning. In A. Abraham, M. Köppen, and K. Franke, editors, *Proc. 3rd Int. Conf. on Hybrid Intelligent Systems*, volume 105 of *Frontiers in Artificial Intelligence and Applications*, pages 399–408. IOS Press, 2003.

[159] A. S. d'Avila Garcez, L. C. Lamb, and D. M. Gabbay. Connectionist computations of intuitionistic reasoning. *Theoretical Computer Science*, 358(1):34–55, 2006.

[160] A. S. d'Avila Garcez, L. C. Lamb, and D. M. Gabbay. Connectionist modal logic: Representing modalities in neural networks. *Theoretical Computer Science*, 371(1-2):34–53, 2007.

[161] A. S. d'Avila Garcez, L. C. Lamb, and D. M. Gabbay. *Neural-Symbolic Cognitive Reasoning*. Cognitive Technologies. Springer, 2009.

[162] A. S. d'Avila Garcez and G. Zaverucha. The connectionist inductive learning and logic programming system. *Applied Intelligence*, 11(1):59–77, 1999.

[163] G. De Cooman and M. Zaffalon. Updating beliefs with incomplete observations. *Artificial Intelligence*, 159(1-2):75–125, 2004.

[164] B. De Finetti. La logique des probabilités. In *Congrès Int. de Philosophie Scientifique*, pages 1–9, Paris, France, 1936. Hermann et Cie.

[165] L. De Raedt. *Logical and Relational Learning*. Springer, 2008.

[166] L. De Raedt, S. Dumancic, R. Manhaeve, and G. Marra. From statistical relational to neuro-symbolic artificial intelligence. In C. Bessiere, editor, *Proc. 29th Int. Joint Conf. on Artificial Intelligence (IJCAI'20)*, pages 4943–4950. ijcai.org, 2020.

[167] L. De Raedt, P. Frasconi, K. Kersting, and S. Muggleton, editors. *Probabilistic Inductive Logic Programming - Theory and Applications*, volume 4911 of *LNCS*. Springer, 2008.

[168] L. De Raedt, T. Guns, and S. Nijssen. Constraint programming for itemset mining. In *Proc. 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, Las Vegas, Aug. 24-27*, pages 204–212, 2008.

[169] L. De Raedt, T. Guns, and S. Nijssen. Constraint programming for data mining and machine learning. In *Proc. 24th AAAI Conf. on Artificial Intelligence, (AAAI'10), Atlanta, July 11-15*, 2010.

[170] L. De Raedt, K. Kersting, S. Natarajan, and D. Poole. *Statistical Relational Artificial Intelligence: Logic, Probability, and Computation*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2016.

[171] L. De Raedt, A. Kimmig, and H. Toivonen. Problog: A probabilistic prolog and its application in link discovery. In *Proc. 20th Int. Joint Conf. on Artificial Intelligence (IJCAI'07), Hyderabad, Jan. 6-12*, pages 2462–2467, 2007.

[172] G. DeJong and R. Mooney. Explanation-based learning: An alternative view. *Machine learning*, 1:145–176, 1986.

[173] E. Demirovic, A. Lukina, E. Hebrard, J. Chan, J. Bailey, C. Leckie, K. Ramamohanarao, and P. J. Stuckey. Murtree: Optimal decision trees via dynamic programming and search. *J. Mach. Learn. Res.*, 23:26:1–26:47, 2022.

[174] A. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[175] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[176] T. Denœux. A *k*-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics*, 25(05):804–813, 1995.

[177] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics A*, 30(2):131–150, 2000.

[178] T. Denœux. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Trans. on Knowledge and Data Engineering*, 25(1):119–130, 2013.

[179] T. Denœux. Logistic regression, neural networks and Dempster-Shafer theory: A new perspective. *Knowl.-Based Syst.*, 176:54–67, 2019.

[180] T. Denœux. Belief functions induced by random fuzzy sets: A general framework for representing uncertain and fuzzy evidence. *Fuzzy Sets and Systems*, 424:63–91, 2021.

[181] T. Denœux. Quantifying Prediction Uncertainty in Regression using Random Fuzzy Sets: the ENNreg model. *IEEE Transactions on Fuzzy Systems*, 31:3690–3699, 2023.

[182] T. Denœux. Reasoning with fuzzy and uncertain evidence using epistemic random fuzzy sets: General framework and practical models. *Fuzzy Sets and Systems*, 453:1–36, 2023.

[183] T. Denœux. Uncertainty quantification in logistic regression using random fuzzy sets and belief functions. *Int. J. of Approx. Reas.*, 168:109159, 2024.

[184] T. Denœux, D. Dubois, and H. Prade. Representations of uncertainty in artificial intelligence: Beyond probability and possibility. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research*, volume 1, chapter 4, pages 119–150. Springer Verlag, 2020.

[185] T. Denœux, D. Dubois, and H. Prade. Representations of uncertainty in artificial intelligence: Probability and possibility. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research*, volume 1, chapter 3, pages 69–117. Springer Verlag, 2020.

[186] T. Denœux, O. Kanjanatarakul, and S. Sriboonchitta. A new evidential k-nearest neighbor rule based on contextual discounting with partially supervised learning. *Int. J. of Approximate Reasoning*, 113:287–302, 2019.

[187] T. Denœux and S. Li. Frequency-calibrated belief functions: Review and new insights. *Int. J. of Approximate Reasoning*, 92:232–254, 2018.

[188] T. Denœux and M.-H. Masson. Evidential reasoning in large partially ordered sets: application to multi-label classification, ensemble clustering and preference aggregation. *Annals of Operations Research*, 195:135–161, 2012.

[189] T. Denœux and L. M. Zouhal. Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets and Systems*, 122(3):47–62, 2001.

[190] V. Derkinderen, R. Manhaeve, P. Zuidberg Dos Martires, and L. De Raedt. Semirings for probabilistic and neuro-symbolic logic programming. *Int. J. of Approximate Reasoning*, this issue:109130, 2024.

[191] J. Derrac and S. Schockaert. Inducing semantic relations from conceptual spaces: a data-driven approach to plausible reasoning. *Artificial Intelligence*, pages 74–105, 2015.

[192] S. Destercke. Uncertain data in learning: challenges and opportunities. In U. Johansson, H. Boström, K. A. Nguyen, Z. Luo, and L. Carlsson, editors, *Proc. 11th Symp. on Conformal and Probabilistic Prediction with Applications, 24-26 Aug. 2022, Brighton*, volume 179 of *Proceedings of Machine Learning Research*, pages 322–332. PMLR, 2022.

[193] M. Diligenti, M. Gori, M. Maggini, and L. Rigutini. Bridging logic and kernel machines. *Machine Learning*, 86(1):57–88, 2012.

[194] B. Dimanov, U. Bhatt, M. Jamnik, and A. Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, and J. Lang, editors, *ECAI 2020 - 24th European conf. on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th conf. on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2473–2480. IOS Press, 2020.

[195] A. Dittadi, T. Bolander, and O. Winther. Learning to plan from raw data in grid-based games. In D. D. Lee, A. Steen, and T. Walsh, editors, *GCAI-2018, 4th Global Conf. on Artificial Intelligence, Luxembourg, September 18-21*, volume 55 of *EPiC Series in Computing*, pages 54–67. EasyChair, 2018.

[196] C. Domshlak, E. Hüllermeier, S. Kaci, and H. Prade. Preferences in AI: An overview. *Artificial Intelligence*, 175(7-8):1037–1052, 2011.

[197] I. Donadello, L. Serafini, and A. S. d'Avila Garcez. Logic tensor networks for semantic image interpretation. In C. Sierra, editor, *Proc. of the 26th Int. Joint Conf. on Artificial Intelligence (IJCAI'17)*, pages 1596–1602. ijcai.org, 2017.

[198] H. Dong, J. Mao, T. Lin, C. Wang, L. Li, and D. Zhou. Neural logic machines. In *Proc. 7th Int. Conf. on Learning Representations (ICLR'19), New Orleans, May 6-9*. OpenReview.net, 2019.

[199] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. https://doi.org/10.48550/arXiv.1702.08608, 2017.

[200] F. K. Dosilovic, M. Brcic, and N. Hlupic. Explainable artificial intelligence: A survey. In *MIPRO*, pages 210–215, 2018.

[201] T. Dreossi, D. J. Fremont, S. Ghosh, E. Kim, H. Ravanbakhsh, M. Vazquez-Chanlatte, and S. A. Seshia. VERIFAI: A toolkit for the design and analysis of artificial intelligence-based systems. *CoRR*, abs/1902.04245, 2019.

[202] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018.

[203] K. Duan, S. S. Keerthi, W. Chu, S. K. Shevade, and A. N. Poo. Multi-category classification by softmax combination of binary classifiers. In T. Windeatt and F. Roli, editors, *Proc. 4th Int. Workshop on Multiple Classifier Systems (MCS 2003)*, volume 2709 of *LNCS*, pages 125–134. Springer, 2003.

[204] D. Dubois. Possibility theory and statistical reasoning. *Comput. Stat. Data Anal.*, 51(1):47–69, 2006.

[205] D. Dubois and E. Hüllermeier. Comparing probability measures using possibility theory: A notion of relative peakedness. *Int. J. Approx. Reasoning*, 45(2):364–385, 2007.

[206] D. Dubois, E. Hüllermeier, and H. Prade. A systematic approach to the assessment of fuzzy association rules. *Data Min. Knowl. Discov.*, 13(2):167–192, 2006.

[207] D. Dubois, J. Lang, and H. Prade. Possibilistic logic. In D. Gabbay, C. Hogger, J. Robinson, and D. Nute, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming, Vol. 3,* pages 439–513. Oxford University Press, 1994.

[208] D. Dubois and H. Prade. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York, 1988.

[209] D. Dubois and H. Prade. What are fuzzy rules and how to use them. *Fuzzy Sets Syst.*, 84(2):169–185, 1996.

[210] D. Dubois and H. Prade. Fuzzy criteria and fuzzy rules in subjective evaluation – A general discussion. In *Proc. 5th Eur. Cong. Intel. Techn. Soft Comput. (EUFIT'97), Aachen, Vol. 1, 975-979*, 1997.

[211] D. Dubois and H. Prade. Possibility theory and formal concept analysis: Characterizing independent sub-contexts. *Fuzzy Sets and Systems*, 196:4–16, 2012.

[212] D. Dubois and H. Prade. Practical methods for constructing possibility distributions. *Int. J. Intell. Syst.*, 31(3):215–239, 2016.

[213] D. Dubois and H. Prade. From possibilistic rule-based systems to machine learning - A discussion paper. In J. Davis and K. Tabia, editors, *Proc. 14th Int. Conf. on Scalable Uncertainty Management (SUM'20), Bozen-Bolzano, Sept.23-25*, volume 12322 of *LNCS*, pages 35–51. Springer, 2020.

[214] D. Dubois and H. Prade. A glance at causality. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research. Vol. 1 : Knowledge Representation, Reasoning and Learning*, pages 275–305. Springer, 2020.

[215] D. Dubois and H. Prade. Reasoning and learning in the setting of possibility theory - overview and perspectives. *Int. J. of Approximate Reasoning*, this issue:109028, 2023.

[216] D. Dubois, H. Prade, and G. Richard. Multiple-valued extensions of analogical proportions. *Fuzzy Sets and Systems*, 292:193–202, 2016.

[217] D. Dubois, H. Prade, and A. Rico. The logical encoding of Sugeno integrals. *Fuzzy Sets and Systems*, 241:61–75, 2014.

[218] D. Dubois, H. Prade, and S. Schockaert. Generalized possibilistic logic: Foundations and applications to qualitative reasoning about uncertainty. *Artif. Intell.*, 252:139–174, 2017.

[219] D. Dubois, H. Prade, and T. Sudkamp. On the representation, measurement, and discovery of fuzzy associations. *IEEE Trans. Fuzzy Systems*, 13(2):250–262, 2005.

[220] K. Dvijotham, M. Garnelo, A. Fawzi, and P. Kohli. Verification of deep probabilistic models. *CoRR*, abs/1812.02795, 2018.

[221] K. Dvijotham, R. Stanforth, S. Gowal, T. A. Mann, and P. Kohli. A dual approach to scalable verification of deep networks. In A. Globerson and R. Silva, editors, *Proc. of the 34th Conf. on Uncertainty in Artificial Intelligence (UAI'18), Monterey, Aug. 6-10*, pages 550–559. AUAI Press, 2018.

[222] K. D. Dvijotham, R. Stanforth, S. Gowal, C. Qin, S. De, and P. Kohli. Efficient neural network verification with exactness characterization. In A. Globerson and R. Silva, editors, *Proc 35th Conf. on Uncertainty in Artificial Intelligence (UAI'19), Tel Aviv, July 22-25*, page 164. AUAI Press, 2019.

[223] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. *CoRR*, abs/1104.3913, 2011.

[224] S. Dzeroski and N. Lavrac, editors. *Relational data mining*. Springer, 2001.

[225] S. Džeroski and B. Ženko. Is combining classifiers with stacking better than selecting the best one? *Machine learning*, 54(3):255–273, 2004.

[226] K. Erk. Representing words as regions in vector space. In *Proc. 13th Conf. on Computational Natural Language Learning*, pages 57–65, 2009.

[227] M. Espinosa Zarlenga, P. Barbiero, G. Ciravegna, G. Marra, F. Giannini, M. Diligenti, Z. Shams, F. Precioso, S. Melacci, A. Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in Neural Information Processing Systems*, 35:21400–21413, 2022.

[228] R. Evans and E. Grefenstette. Learning explanatory rules from noisy data. *J. Artif. Intell. Res.*, 61:1–64, 2018.

[229] R. Evans and E. Grefenstette. Learning explanatory rules from noisy data (extended abstract). In Lang [392], pages 5598–5602.

[230] R. Evans, D. Saxton, D. Amos, P. Kohli, and E. Grefenstette. Can neural networks understand logical entailment? In *Proc. 6th Int. Conf. on Learning Representations (ICLR'18), Vancouver, Apr. 30 - May 3*, 2018.

[231] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. In *Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'18), Salt Lake City, June 18-22*, pages 1625–1634. IEEE Computer Society, 2018.

[232] M. A. Fahandar and E. Hüllermeier. Learning to rank based on analogical reasoning. In *Proc. 32th Nat Conf. on Artificial Intelligence (AAAI'18), New Orleans, Feb. 2-7*, 2018.

[233] M. A. Fahandar, E. Hüllermeier, and I. Couso. Statistical inference for incomplete ranking data: the case of rank-dependent coarsening. In *Int. Conf. on Machine Learning*, pages 1078–1087. PMLR, 2017.

[234] S. Fakhraei, L. Raschid, and L. Getoor. Drug-target interaction prediction for drug repurposing with probabilistic similarity logic, 2013.

[235] H. Fargier, S. Mengel, and J. Mengin. An extended knowledge compilation map for conditional preference statements-based and generalized additive utilities-based languages. *Annals of Mathematics and Artificial Intelligence*, To appear, 2024.

[236] H. Fargier and J. Mengin. A knowledge compilation map for conditional preference statements-based languages. In F. Dignum, A. Lomuscio, U. Endriss, and A. Nowé, editors, *Proc. 20th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS '21)*, page 492–500. ACM, 2021.

[237] G. Farnadi, S. H. Bach, M. Moens, L. Getoor, and M. D. Cock. Extending PSL with fuzzy quantifiers. In *Statistical Relational Artificial Intelligence, Papers from the 2014 AAAI Workshop, Québec City, Québec, Canada, July 27, 2014*, volume WS-14-13 of *AAAI Workshops*, pages 35–37, 2014.

[238] H. Farreny and H. Prade. Default and inexact reasoning with possibility degrees. *IEEE Trans. on Syst., Man, and Cyber.*, 16(2):270–276, 1986.

[239] H. Farreny and H. Prade. Positive and negative explanations of uncertain reasoning in the framework of possibility theory. In *Proc. 5th conf. on Uncertainty in Artificial Intelligence (UAI'89), Windsor, ON, Aug. 18-20*, pages 95–101, 1989. Available in CoRR, abs/1304.1502, 2013; Expanded version: Explications de raisonnements dans l'incertain), Revue d'Intelligence Artificielle, 4(2), 43-75,1990.

[240] H. Farreny and H. Prade. *Positive and Negative Explanations of Uncertain Reasoning in the Framework of Possibility Theory*, page 319–333. John Wiley & Sons, Inc., USA, 1992.

[241] T. Fel, I. Felipe, D. Linsley, and T. Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[242] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proc. 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA, 2015. Association for Computing Machinery.

[243] S. Ferré and O. Ridoux. Introduction to logical information systems. *Information Process. & Manag.*, 40(3):383–419, 2004.

[244] S. Ferré, M. Kaytoue, M. Huchard, S. O. Kuznetsov, and A. Napoli. Formal concept analysis: From knowledge discovery to knowledge processing. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research. Vol. 2 Artificial Intelligence Algorithms*, pages 411–445. Springer-Verlag, 2020.

[245] D. Fierens, H. Blockeel, M. Bruynooghe, and J. Ramon. Logical bayesian networks and their relation to other probabilistic logical models. In *Inductive Logic Programming, Proc. 15th Int. Conf. ILP'05, Bonn, Germany, Aug. 10-13*, pages 121–135, 2005.

[246] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

[247] M. Fischer, M. Balunovic, D. Drachsler-Cohen, T. Gehr, C. Zhang, and M. T. Vechev. DL2: training and querying neural networks with logic. In *Proc 36th Int. Conf. on Machine Learning (ICML'19), Long Beach June 9-15*, pages 1931–1941, 2019.

[248] P. C. Fishburn. Interdependence and additivity in multivariate, unidimensional expected utility theory. *Int. Economic Review*, 8(3):pp. 335–342, 1967.

[249] A. Flint and M. B. Blaschko. Perceptron learning of SAT. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conf. on Neural Information Processing Systems 2012. Proc. of a meeting held Dec. 3-6, 2012, Lake Tahoe, Nevada*, pages 2780–2788, 2012.

[250] M. V. M. França, G. Zaverucha, and A. S. d'Avila Garcez. Fast relational learning using bottom clause propositionalization with artificial neural networks. *Machine Learning*, 94(1):81–104, 2014.

[251] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proc. of the conf. on Fairness, Accountability, and Transparency*, FAT* '19, page 329–338, New York, NY, USA, 2019. Association for Computing Machinery.

[252] J. Fürnkranz and E. Hüllermeier, editors. *Preference Learning*. Springer, 2010.

[253] J. Fürnkranz, E. Hüllermeier, C. Rudin, R. Slowinski, and S. Sanner. Preference learning (dagstuhl seminar 14101). *Dagstuhl Reports*, 4(3):1–27, 2014.

[254] G. P. G. Leng, Th. M. McGinnity. An approach for on-line extraction of fuzzy rules using a self-organising fuzzy neural network. *Fuzzy Sets and Systems*, 150:211—-243, 2005.

[255] M. H. Gad-Elrab, D. Stepanova, J. Urbani, and G. Weikum. Exception-enriched rule learning from knowledge graphs. In P. Groth, E. Simperl, A. J. G. Gray, M. Sabou, M. Krötzsch, F. Lécué, F. Flöck, and Y. Gil, editors, *Proc. 15th Int. Semantic Web Conf.(ISWC'16), Part I Kobe, Oct. 17-21*, volume 9981 of *LNCS*, pages 234–251, 2016.

[256] L. A. Galárraga, C. Teflioudi, K. Hose, and F. M. Suchanek. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In D. Schwabe, V. A. F. Almeida, H. Glaser, R. Baeza-Yates, and S. B. Moon, editors, *Proc. 22nd Int. World Wide Web Conf., WWW '13, Rio de Janeiro, May 13-17*, pages 413–422. Int. World Wide Web Conf. Steering Committee / ACM, 2013.

[257] A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *Proc. 14th Conf. on Uncertainty in AI*, pages 148–155. Morgan Kaufmann, 1998.

[258] B. Ganter and S. O. Kuznetsov. Pattern structures and their projections. In H. S. Delugach and G. Stumme, editors, *Proc. 9th Int. Conf. on Conceptual Structures (ICCS'01)*, volume 2120 of *LNCS*, pages 129–142. Springer, 2001.

[259] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, 1998.

[260] A. S. d. Garcez, K. B. Broda, and D. M. Gabbay. *Neural-symbolic learning systems: foundations and applications*. Springer Science & Business Media, 2002.

[261] P. Gärdenfors. Nonmonotonic inference, expectations, and neural networks. In R. Kruse and P. Siegel, editors, *Proc. Europ. Conf. on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQA), Marseille, Oct. 15-17*, volume 548 of *LNCS*, pages 12–27. Springer, 1991.

[262] P. Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2000.

[263] T. Gehr, M. Mirman, D. Drachsler-Cohen, P. Tsankov, S. Chaudhuri, and M. T. Vechev. AI2: safety and robustness certification of neural networks with abstract interpretation. In *Proc. 2018 IEEE Symp. on Security and Privacy (SP'18), May 21-23, San Francisco*, pages 3–18. IEEE Computer Society, 2018.

[264] L. Getoor and B. Taskar, editors. *Introduction to Statistical Relational Learning*. Adaptive Computation and Machine Learning. MIT Press, 2007.

[265] B. Ghosh and K. S. Meel. IMLI: an incremental framework for maxsat-based learning of interpretable classification rules. In V. Conitzer, G. K. Hadfield, and S. Vallor, editors, *Proc. of the 2019 AAAI/ACM Conf. on AI, Ethics, and Society, AIES 2019, Honolulu, January 27-28*, pages 203–210. ACM, 2019.

[266] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In D. Precup and Y. W. Teh, editors, *Proc. of the 34th Int. Conf. on Machine Learning (ICML'17), Sydney, Aug. 6-11*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 2017.

[267] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In F. Bonchi, F. J. Provost, T. Eliassi-Rad, W. Wang, C. Cattuto, and R.Ghani, editors, *Proc. 5th IEEE Int. Conf. on Data Science and Advanced Analytics (DSAA'18), Turin, Oct. 1-3*, pages 80–89. IEEE, 2018.

[268] L. H. Gilpin, C. Testart, N. Fruchter, and J. Adebayo. Explaining explanations to society. *CoRR*, abs/1901.06560, 2019.

[269] E. Giunchiglia, A. Tatomir, M. C. Stoian, and T. Lukasiewicz. CCN+: A neuro-symbolic framework for deep learning with requirements. *Int. J. of Approximate Reasoning*, this issue:109124, 2024.

[270] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. of Computational and Graphical Statistics*, 2015.

[271] I. J. Goodfellow, M. Mirza, X. Da, A. C. Courville, and Y. Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. In *Conf. Track Proc. 2nd Int. Conf. on Learning Representations (ICLR'14), Banff, April 14-16*, 2014.

[272] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples, 2014.

[273] D. Gopinath, G. Katz, C. S. Pasareanu, and C. W. Barrett. Deepsafe: A data-driven approach for assessing robustness of neural networks. In Lahiri and Wang [389], pages 3–19.

[274] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee. Counterfactual visual explanations. In *Proc. 36th Int. Conf. on Machine Learning (ICML)*, 2019.

[275] M. Grabisch and C. Labreuche. A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. *Annals of Oper. Res.*, 175:247–286, 2010.

[276] Y. Grandvalet, A. Rakotomamonjy, J. Keshet, and S. Canu. Support vector machines with a reject option. In *Advances in neural information processing systems*, pages 537–544, 2009.

[277] S. Greco, M. Inuiguchi, and R. Slowinski. Fuzzy rough sets and multiple-premise gradual decision rules. *Int. J. Approx. Reason.*, 41(2):179–211, 2006.

[278] S. Greco, M. Inuiguchi, and R. Slowinski. Fuzzy rough sets and multiple-premise gradual decision rules. *Int. J. Approx. Reasoning*, 41(2):179–211, 2006.

[279] S. Greco, B. Matarazzo, and R. Slowinski. Axiomatic characterization of a general utility function and its particular cases in terms of conjoint measurement and rough-set decision rules. *Europ.J. of Operational Research*, 158(2):271–292, 2004.

[280] C. Grozea and M. Popescu. Can machine learning learn a decision oracle for NP problems? A test on SAT. *Fundam. Inform.*, 131(3-4):441–450, 2014.

[281] J. W. Grzymala-Busse. LERS - A data mining system. In O. Maimon and L. Rokach, editors, *The Data Mining and Knowledge Discovery Handbook*, pages 1347–1351. Springer, 2005.

[282] J. W. Grzymala-Busse. Rough set theory with applications to data mining. In *Real World Applications of Computational Intelligence*, pages 221–244. Springer, 2005.

[283] J. W. Grzymala-Busse and Y. Yao. Probabilistic rule induction with the LERS data mining system. *Int. J. Intell. Syst.*, 26(6):518–539, 2011.

[284] J. W. Grzymala-Busse and W. Ziarko. Data mining and rough set theory. *Commun. ACM*, 43(4):108–109, 2000.

[285] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling. Scene graph generation with external knowledge and image reconstruction. In *IEEE conf. on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1969–1978. Computer Vision Foundation / IEEE, 2019.

[286] R. Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 2022.

[287] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1–42, 2018.

[288] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, 2019.

[289] J. L. Guigues and V. Duquenne. Familles minimales d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences Humaines*, 95:5–18, 1986.

[290] R. Guillaume and D. Dubois. A maximum likelihood approach to inference under coarse data based on minimax regret. In S. Destercke, T. Denœux, M. A. Gil, P. Grzegorzewski, and O. Hryniewicz, editors, *Uncertainty Modelling in Data Science, SMPS 2018*, volume 832 of *Advances in Intelligent Systems and Computing*, pages 99–106. Springer, 2018.

[291] D. Gunning, E. Vorm, J. Y. Wang, and M. Turek. DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 2(4), 2021.

[292] A. Gupta, G. Boleda, and S. Padó. Instantiation. *CoRR*, abs/1808.01662, 2018.

[293] V. Gutiérrez-Basulto and S. Schockaert. From knowledge graph embedding to ontology embedding? an analysis of the compatibility between vector space representations and rules. In *Proc. of the 16th Int. Conf. on Principles of Knowledge Representation and Reasoning*, pages 379–388, 2018.

[294] I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors. *Advances in Neural Information Processing Systems 30: Annual Conf. on Neural Information Processing Systems 2017, Dec. 4-9, Long Beach*, 2017.

[295] T. M. Ha. The optimum class-selective rejection rule. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(6):608–615, 1997.

[296] S. Haim and T. Walsh. Restart strategy selection using machine learning techniques. In O. Kullmann, editor, *Proc. 12th Int. Conf. on Theory and Applications of Satisfiability Testing (SAT'09), Swansea, June 30 - July 3*, volume 5584 of *LNCS*, pages 312–325. Springer, 2009.

[297] P. Hájek and P. Havránek. *Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory*. Springer Verlag, 1978.

[298] J. Y. Halpern. *Reasoning about Uncertainty*. MIT Press, second edition, 2017. First edition published in 2005.

[299] J. Y. Halpern, R. Fagin, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge*. MIT Press, 1995 & 2003.

[300] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach. part II: explanations. *The British Journal for the Philosophy of Science*, 56(4):889–911, 2005.

[301] D. W. Hasling, W. J. Clancey, and G. Rennels. Strategic explanations for a diagnostic consultation system. *Int. J. of Man-Machine Studies*, 20(1):3–19, 1984.

[302] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. Neural collaborative filtering. In *Proc. 26th Int. Conf. on World Wide Web*, pages 173–182, 2017.

[303] D. Heaven. Why deep-learning AIs are so easy to fool. *Nature*, 574(7777):163, 2019.

[304] S. Heid, J. Hanselle, J. Fürnkranz, and E. Hüllermeier. Learning decision catalogues for situated decision making: The case of scoring systems. *Int. J. of Approximate Reasoning*, this issue:????, 2024.

[305] D. F. Heitjan and D. B. Rubin. Ignorability and coarse data. *Ann. Statist.*, 19(4):2244–2253, 1991.

[306] J. Heo, S. Joo, and T. Moon. Fooling neural network interpretations via adversarial model manipulation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[307] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *Int. Conf. on Learning Representations*, volume 3, 2017.

[308] F. Hill, K. Cho, A. Korhonen, and Y. Bengio. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30, 2016.

[309] P. Hitzler, R. Rayan, J. Zalewski, S. S. Norouzi, A. Eberhart, and E. Y. Vasserman. Deep deductive reasoning is a hard deep learning problem. *Neurosymbolic Artificial Intelligence*, under review.

[310] P. Hitzler and M. K. Sarker, editors. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, volume 342 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2021.

[311] Y. Hmidy, A. Rico, and O. Strauss. Macsum aggregation learning. *Fuzzy Sets and Systems*, 459:182–200, 2023.

[312] R. R. Hoffman and G. Klein. Explaining explanation, part 1: Theoretical foundations. *IEEE Intelligent Systems*, 32(3):68–73, 2017.

[313] R. R. Hoffman, T. Miller, S. T. Mueller, G. Klein, and W. J. Clancey. Explaining explanation, part 4: A deep dive on deep nets. *IEEE Intelligent Systems*, 33(3):87–95, 2018.

[314] R. R. Hoffman, S. T. Mueller, and G. Klein. Explaining explanation, part 2: Empirical foundations. *IEEE Intelligent Systems*, 32(4):78–86, 2017.

[315] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable AI: challenges and prospects. *CoRR*, abs/1812.04608, 2018.

[316] P. Hohenecker and T. Lukasiewicz. Deep learning for ontology reasoning. *CoRR*, abs/1705.10342, 2017.

[317] S. Hölldobler, Y. Kalinke, and H. Störr. Approximating the semantics of logic programs by recurrent neural networks. *Applied Intelligence*, 11(1):45–58, 1999.

[318] J. N. Hooker, editor. *Proc. 24th Int. Conf. on Principles and Practice of Constraint Programming (CP'18), Lille, Aug. 27-31*, volume 11008 of *LNCS*. Springer, 2018.

[319] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[320] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *J. of Machine Learning Research*, 5:1457–1469, 2004.

[321] X. Hu, C. Rudin, and M. I. Seltzer. Optimal sparse decision trees. In *NeurIPS*, 2019.

[322] Y. Hu, A. Chapman, G. Wen, and W. Hall. What can knowledge bring to machine learning? - A survey of low-shot learning for structured data. *ACM Transactions on Intelligent Systems and Technology*, 13(3):48:1–48:45, 2022.

[323] Z. Hu, X. Ma, Z. Liu, E. H. Hovy, and E. P. Xing. Harnessing deep neural networks with logic rules. In *Proc. 54th Annual Meeting of the Association for Computational Linguistics*, 2016.

[324] D. Huang. On learning to prove. *CoRR*, abs/1904.11099, 2019.

[325] D. Huang, P. Dhariwal, D. Song, and I. Sutskever. Gamepad: A learning environment for theorem proving. In *Proc. 7th Int. Conf. on Learning Representations (ICLR'19), New Orleans, May 6-9*. OpenReview.net, 2019.

[326] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu. Safety verification of deep neural networks. In Majumdar and Kuncak [429], pages 3–29.

[327] X. Huang and J. Marques-Silva. From robustness to explainability and back again. *CoRR*, abs/2306.03048, 2023.

[328] X. Huang and J. Marques-Silva. On the failings of Shapley values for explainability. *Int. J. of Approximate Reasoning*, this issue:109112, 2024.

[329] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. Binarized neural networks. In Lee et al. [399], pages 4107–4115.

[330] E. Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *Int. J. of Approx. Reas.*, 55(7):1519–1534, 2014.

[331] E. Hüllermeier and W. Cheng. Superset learning based on generalized loss minimization. In *Machine Learning and Knowledge Discovery in Databases - Proc. Eur. Conf., ECML PKDD 2015, Part II*, volume 9285 of *LNCS*, pages 260–275. Springer, 2015.

[332] E. Hüllermeier, S. Destercke, and I. Couso. Learning from imprecise data: Adjustments of optimistic and pessimistic variants. In *Proc. 13th Int. conf. on Scalable Uncertainty Management (SUM'19), Compiègne, Dec. 16-18*, pages 266–279, 2019.

[333] E. Hüllermeier, D. Dubois, and H. Prade. Model adaptation in possibilistic instance-based reasoning. *IEEE Trans. Fuzzy Systems*, 10(3):333–339, 2002.

[334] E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[335] F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In C. A. C. Coello, editor, *Proc. 5th Int. Conf. on Learning and Intelligent Optimization - , LION 5, Rome, January 17-21. Selected Papers*, volume 6683 of *LNCS*, pages 507–523. Springer, 2011.

[336] F. Hutter, H. H. Hoos, K. Leyton-Brown, and T. Stützle. ParamILS: An automatic algorithm configuration framework. *J. Artif. Intell. Res.*, 36:267–306, 2009.

[337] F. Hutter, L. Xu, H. H. Hoos, and K. Leyton-Brown. Algorithm runtime prediction: Methods & evaluation. *Artif. Intell.*, 206:79–111, 2014.

[338] L. Hyafil and R. L. Rivest. Constructing optimal binary decision trees is np-complete. *Information Processing Letters*, 5(1):15–17, 1976.

[339] ICLR. *Proc. 5th Int. Conf. on Learning Representations (ICLR'17), Toulon, Apr. 24-26*. OpenReview.net, 2017.

[340] A. Ignatiev, N. Narodytska, N. Asher, and J. Marques-Silva. From contrastive to abductive explanations and back again. In M. Baldoni and S. Bandini, editors, *AIxIA 2020 - Advances in Artificial Intelligence - XIXth International Conference of the Italian Association for Artificial Intelligence, Virtual Event, November 25-27, 2020, Revised Selected Papers*, volume 12414 of *LNCS*, pages 335–355. Springer, 2020.

[341] A. Ignatiev, N. Narodytska, and J. Marques-Silva. Abduction-based explanations for machine learning models. In *Proc. of the 33rd Int. Conf. on Artificial Intelligence, AAAI19*, pages 1511–1519, 2019.

[342] A. Ignatiev, N. Narodytska, and J. Marques-Silva. Abduction-based explanations for machine learning models. In *Proc. 33rd AAAI Conf. on Artificial Intelligence (AAAI'19), The 31st Innovative Applications of Artificial Intelligence Conf., IAAI 2019, The 9th AAAI Symp. on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Jan. 27 - Feb. 1* [1], pages 1511–1519.

[343] A. Ignatiev, N. Narodytska, and J. Marques-Silva. On validating, repairing and refining heuristic ML explanations. *CoRR*, abs/1907.02509, 2019.

[344] A. Ignatiev, F. Pereira, N. Narodytska, and J. Marques-Silva. A SAT-based approach to learn explainable decision sets. In D. Galmiche, S. Schulz, and R. Sebastiani, editors, *Proc. 9th Int. Joint Conf. Automated Reasoning (IJCAR'18), held as Part of the Federated Logic Conference, FloC 2018, Oxford, UK, July 14-17*, volume 10900 of *LNCS*, pages 627–645. Springer, 2018.

[345] G. Irving, C. Szegedy, A. A. Alemi, N. Eén, F. Chollet, and J. Urban. DeepMath - deep sequence models for premise selection. In Lee et al. [399], pages 2235–2243.

[346] Y. Izza, A. Ignatiev, and J. Marques-Silva. On explaining decision trees. *CoRR*, abs/2010.11034, 2020.

[347] Y. Izza, A. Ignatiev, and J. Marques-Silva. On tackling explanation redundancy in decision trees. *J. Artif. Intell. Res.*, 75:261–321, 2022.

[348] S. Jabbour, L. Sais, and Y. Salhi. Mining top-k motifs with a sat-based framework. *Artif. Intell.*, 244:30–47, 2017.

[349] M. Jaeger. Ignorability in statistical and probabilistic inference. *J. Artif. Intell. Res.*, 24:889–917, 2005.

[350] S. Jameel and S. Schockaert. Entity embeddings with conceptual subspaces as a basis for plausible reasoning. In *Proc. 22nd Europ. Conf. on Artificial Intelligence (ECAI'16), 29 Aug.-2 Sept. 2016, The Hague*, pages 1353–1361, 2016.

[351] S. Jameel and S. Schockaert. Modeling context words as regions: An ordinal regression approach to word embedding. In *Proc. 21st Conf. on Computational Natural Language Learning*, pages 123–133, 2017.

[352] J. Jang and C. Sun. Functional equivalence between radial basis function networks and fuzzy inference systems. *IEEE Trans. Neural Networks*, 4(1):156–159, 1993.

[353] M. Janota. Towards generalization in QBF solving via machine learning. In *Proc. 32nd AAAI Conf. on Artificial Intelligence, (AAAI-18), New Orleans, Feb. 2-7*, pages 6607–6614, 2018.

[354] M. Janota and I. Lynce, editors. *Proc. 22nd Int. Conf. on Theory and Applications of Satisfiability Testing - SAT'19, Lisbon, July 9-12*, volume 11628 of *LNCS*. Springer, 2019.

[355] R. Jeffrey. *The logic of decision*. 2nd ed. Chicago University Press, 1983.

[356] P. Jung, G. Marra, and O. Kuželka. Quantified neural Markov logic networks. *Int. J. of Approximate Reasoning*, this issue:109172, 2024.

[357] U. Junker, J. Delgrande, J. Doyle, F. Rossi, and T. Schaub. Preface to the special issue of Computational Intelligence on preferences. *Computational Intelligence*, 20(2):109–110, 2004.

[358] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[359] C. Kaliszyk, F. Chollet, and C. Szegedy. Holstep: A machine learning dataset for higher-order logic theorem proving. In *Proc. 5th Int. Conf. on Learning Representations (ICLR'17), Toulon, Apr. 24-26* [339].

[360] C. Kaliszyk and J. Urban. Learning-assisted automated reasoning with Flyspeck. *J. Autom. Reasoning*, 53(2):173–213, 2014.

[361] C. Kaliszyk and J. Urban. Learning-assisted theorem proving with millions of lemmas. *J. Symb. Comput.*, 69:109–128, 2015.

[362] C. Kaliszyk, J. Urban, H. Michalewski, and M. Olsák. Reinforcement learning of theorem proving. In Bengio et al. [60], pages 8836–8847.

[363] C. Kaliszyk, J. Urban, and J. Vyskocil. Machine learner for automated reasoning 0.4 and 0.5. In S. Schulz, L. de Moura, and B. Konev, editors, *4th Workshop on Practical Aspects of Automated Reasoning, PAAR@IJCAR 2014, Vienna, 2014*, volume 31 of *EPiC Series in Computing*, pages 60–66. EasyChair, 2014.

[364] O. Kanjanatarakul, S. Sriboonchitta, and T. Denœux. Statistical estimation and prediction using belief functions: principles and application to some econometric models. *Int. J. of Approximate Reasoning*, 72:71–94, 2016.

[365] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2023.

[366] A.-H. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proc. of the 2021 ACM Conf. on Fairness, Accountability, and Transparency, FAccT'21*, pages 353–362, 2021.

[367] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. of the IEEE conf. on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[368] G. Kassel. The use of deep knowledge to improve explanation capabilities of rule-based expert systems. In H. Balzert, G. Heyer, and R. Lutze, editors, *Expertensysteme '87: Konzepte und Werkzeuge, Tagung I/1987 des German Chapter of the ACM am 7. und 8.4.1987 in Nürnberg*, volume 28 of *Berichte des German Chapter of the ACM*, pages 315–326. Teubner, 1987.

[369] G. Katz, C. W. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In Majumdar and Kuncak [429], pages 97–117.

[370] G. Katz, D. A. Huang, D. Ibeling, K. Julian, C. Lazarus, R. Lim, P. Shah, S. Thakoor, H. Wu, A. Zeljic, D. L. Dill, M. J. Kochenderfer, and C. W. Barrett. The marabou framework for verification and analysis of deep neural networks. In I. Dillig and S. Tasiran, editors, *Proc. 31st Int. Conf. on Computer Aided Verification (CAV'19), New York City, July 15-18, Part I*, volume 11561 of *LNCS*, pages 443–452. Springer, 2019.

[371] S. M. Kazemi and D. Poole. Simple embedding for link prediction in knowledge graphs. In *Advances in Neural Information Processing Systems 31: Annual Conf. on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 4289–4300, 2018.

[372] M. T. Keane, E. M. Kenny, E. Delaney, and B. Smyth. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. In *Proc. of the 30th Int. Conf. on Artificial Intelligence, IJCAI21*, pages 4466–4474, 2021.

[373] A. Kemmar, Y. Lebbah, S. Loudni, P. Boizumault, and T. Charnois. Prefix-projection global constraint and top-k approach for sequential pattern mining. *Constraints*, 22(2):265–306, 2017.

[374] E. B. Khalil, H. Dai, Y. Zhang, B. Dilkina, and L. Song. Learning combinatorial optimization algorithms over graphs. In Guyon et al. [294], pages 6348–6358.

[375] M. Khiari, P. Boizumault, and B. Crémilleux. Constraint programming for mining n-ary patterns. In *Principles and Practice of Constraint Programming - Proc. CP 16th Int. Conf. CP 2010, St. Andrews, Scotland, Sept. 6-10*, pages 552–567, 2010.

[376] A. R. KhudaBukhsh, L. Xu, H. H. Hoos, and K. Leyton-Brown. SATenstein: automatically building local search SAT solvers from components. *Artif. Intell.*, 232:20–42, 2016.

[377] G. Klein. Explaining explanation, part 3: The causal landscape. *IEEE Intelligent Systems*, 33(2):83–88, 2018.

[378] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. Concept bottleneck models. In *Int. Conf. on Machine Learning*, pages 5338–5348. PMLR, 2020.

[379] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[380] D. Kreiss, G. Schollmeyer, and T. Augustin. Towards improving electoral forecasting by including undecided voters and interval-valued prior knowledge. In *Int. Symp. on Imprecise Probability: Theories and Applications (ISIPTA)*, pages 201–209. PMLR, 2021.

[381] B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, and R. Rastogi, editors. *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Francisco, Aug. 13-17*. ACM, 2016.

[382] C. Kuo, S. S. Ravi, T. Dao, C. Vrain, and I. Davidson. A framework for minimal clustering modification via constraint programming. In *Proc. 31st AAAI Conf. on Artificial Intelligence, February 4-9, 2017, San Francisco*, pages 1389–1395, 2017.

[383] O. Kuzelka, J. Davis, and S. Schockaert. Encoding Markov logic networks in possibilistic logic. In M. Meila and T. Heskes, editors, *Proc. 31st Conf. on Uncertainty in Artificial Intelligence (UAI'15), July 12-16, Amsterdam*, pages 454–463. AUAI Press, 2015.

[384] O. Kuzelka, J. Davis, and S. Schockaert. Learning possibilistic logic theories from default rules. In S. Kambhampati, editor, *Proc. 25th Int. Joint Conf. on Artificial Intelligence (IJCAI'16), New York, 9-15 July*, pages 1167–1173. IJCAI/AAAI Press, 2016.

[385] O. Kuzelka, J. Davis, and S. Schockaert. Induction of interpretable possibilistic logic theories from relational data. In C. Sierra, editor, *Proc. 26th Int. Joint Conf. on Artificial Intelligence (IJCAI'17), Melbourne, Aug. 19-25*, pages 1153–1159. ijcai.org, 2017.

[386] M. Z. Kwiatkowska. Safety verification for deep neural networks with provable guarantees (invited paper). In W. Fokkink and R. van Glabbeek, editors, *Proc. 30th Int. Conf. on Concurrency Theory (CONCUR'19), Aug. 27-30, Amsterdam*, volume 140 of *LIPIcs*, pages 1:1–1:5. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.

[387] C. Labreuche. A general framework for explaining the results of a multi-attribute preference model. *Artif. Intell.*, 175(7-8):1410–1448, 2011.

[388] N. Lachiche and P. A. Flach. 1bc2: A true first-order bayesian classifier. In *12th Int. Conf.on Inductive Logic Programming (ILP'02), Sydney July 9-11, Revised Papers*, pages 133–148, 2002.

[389] S. K. Lahiri and C. Wang, editors. *Proc. 16th Int. Symp. on Automated Technology for Verification and Analysis (ATVA'18), Los Angeles, Oct. 7-10*, volume 11138 of *LNCS*. Springer, 2018.

[390] H. Lakkaraju, S. H. Bach, and J. Leskovec. Interpretable decision sets: A joint framework for description and prediction. In Krishnapuram et al. [381], pages 1675–1684.

[391] A. Lallouet, M. Lopez, L. Martin, and C. Vrain. On learning constraint problems. In *Proc. 22nd IEEE Int. Conf. on Tools with Artificial Intelligence (ICTAI'10), Arras, Oct.27-29, Vol. 1*, pages 45–52. IEEE Computer Society, 2010.

[392] J. Lang, editor. *Proc. of the 27th Int. Joint Conf. on Artificial Intelligence (IJCAI'18), Stockholm, July 13-19*. ijcai.org, 2018.

[393] N. Lavrac, S. Dzeroski, and M. Grobelnik. Learning nonrecursive definitions of relations with LINUS. In *Machine Learning - EWSL-91, European Working Session on Learning, Porto, Portugal, March 6-8, 1991, Proc.*, pages 265–281, 1991.

[394] M. Law, N. Thome, and M. Cord. Learning a distance metric from relative comparisons between quadruplets of images. *Int. J. of Computer Vision*, 121(1):65–94, 2017.

[395] Y. Le Cun. *Quand la Machine Apprend. La révolution des neurones artificiels et de l'apprentissage profond*. Odile Jacob, 2019.

[396] F. Lécué, J. Chen, J. Z. Pan, and H. Chen. Augmenting transfer learning with semantic reasoning. In S. Kraus, editor, *Proc. 28th Int. Joint Conf. on Artificial Intelligence (IJCAI'19), Macao, Aug. 10-16*, pages 1779–1785. ijcai.org, 2019.

[397] G. Lederman, M. N. Rabe, and S. A. Seshia. Learning heuristics for automated reasoning through deep reinforcement learning. *CoRR*, abs/1807.08058, 2018.

[398] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[399] D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors. *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016.

[400] F. Leofante, N. Narodytska, L. Pulina, and A. Tacchella. Automated verification of neural networks: Advances, challenges and perspectives. *CoRR*, abs/1805.09938, 2018.

[401] H. Levesque. A fundamental tradeoff in knowledge representation and reasoning (revised version). In R. Brachman and H. Levesque, editors, *in Readings in Knowledge Representation*, pages 41–70. Morgan Kaufman, 1985.

[402] H. J. Levesque. Knowledge representation and reasoning. *Annual Review of Computer Science*, 1(1):255–288, 1986.

[403] H. J. Levesque and R. J. Brachman. Expressiveness and tractability in knowledge representation and reasoning. *Computational Intelligence*, 3:78–93, 1987.

[404] O. Li, H. Liu, C. Chen, and C. Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In S. A. McIlraith and K. Q. Weinberger, editors, *Proc. of the 32nd AAAI Conf. on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symp. on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Feb. 2-7*, pages 3530–3537. AAAI Press, 2018.

[405] Z. Li, Q. Chen, and V. Koltun. Combinatorial optimization with graph convolutional networks and guided tree search. In Bengio et al. [60], pages 537–546.

[406] C. Lian, S. Ruan, and T. Denœux. Dissimilarity metric learning in the belief function framework. *IEEE Transactions on Fuzzy Systems*, 24(6):1555–1564, 2016.

[407] J. H. Liang, V. Ganesh, P. Poupart, and K. Czarnecki. Exponential recency weighted average branching heuristic for SAT solvers. In D. Schuurmans and M. P. Wellman, editors, *Proc. 30th AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix*, pages 3434–3440. AAAI Press, 2016.

[408] J. H. Liang, V. Ganesh, P. Poupart, and K. Czarnecki. Learning rate based branching heuristic for SAT solvers. In N. Creignou and D. L. Berre, editors, *Theory and Applications of Satisfiability Testing - SAT 2016 - Proc. 19th Int. Conf., Bordeaux, July 5-8*, volume 9710 of *LNCS*, pages 123–140. Springer, 2016.

[409] J. H. Liang, C. Oh, M. Mathew, C. Thomas, C. Li, and V. Ganesh. Machine learning-based restart policy for CDCL SAT solvers. In O. Beyersdorff and C. M. Wintersteiger, editors, *Proc. 21st Int. Conf. SAT'18 on Theory and Applications of Satisfiability Testing (SAT'18), held as Part of the Federated Logic Conference, FloC 2018, Oxford, July 9-12*, volume 10929 of *LNCS*, pages 94–110. Springer, 2018.

[410] Q. V. Liao and K. R. Varshney. Human-centered explainable AI (XAI): From algorithms to user experiences. https://doi.org/10.48550/arXiv.2110.10790, 2022.

[411] J. Lieber, E. Nauer, H. Prade, and G. Richard. Making the best of cases by approximation, interpolation and extrapolation. In M. T. Cox, P. Funk, and S. Begum, editors, *Proc. 26th Int. Conf. on Case-Based Reasoning (ICCBR'18), Stockholm, July 9-12*, volume 11156 of *LNCS*, pages 580–596. Springer, 2018.

[412] J. Lienen and E. Hüllermeier. Credal self-supervised learning. *Advances in Neural Information Processing Systems*, 34:14370–14382, 2021.

[413] J. Lienen and E. Hüllermeier. From label smoothing to label relaxation. In *Proc. 35th AAAI conf. on artificial intelligence (AAAI'21), Virtual Event, Feb. 2-9*, pages 8583–8591, 2021.

[414] J. Lienen and E. Hüllermeier. Mitigating label noise through data ambiguation. In *Proc. 38th AAAI Conf. on Artificial Intelligence (AAAI'24), Vancouver, Feb. 20-27*, pages 13799–13807, 2024.

[415] X. Lin, H. Zhu, R. Samanta, and S. Jagannathan. ART: abstraction refinement-guided training for provably correct neural networks. *CoRR*, abs/1907.10662, 2019.

[416] D. Linsley, S. Eberhardt, T. Sharma, P. Gupta, and T. Serre. What are the visual features underlying human versus machine vision? In *Proc. IEEE Int. Conf. on Computer Vision Workshops*, pages 2706–2714, 2017.

[417] Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[418] R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*, volume 793. Wiley, 2019.

[419] L. Liu and T. Dietterich. Learnability of the superset label learning problem. In *Int. Conf. on Machine Learning*, pages 1629–1637, 2014.

[420] N. Liu, H. Yang, and X. Hu. Adversarial detection with model interpretation. In Y. Guo and F. Farooq, editors, *Proc. of the 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining (KDD'18), London, August 19-23*, pages 1803–1811. ACM, 2018.

[421] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang. K-BERT: enabling language representation with knowledge graph. In *The Thirty-Fourth AAAI conf. on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence conf., IAAI 2020, The Tenth AAAI Symp. on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2901–2908. AAAI Press, 2020.

[422] F. Locatello, S. Bauer, M. Lucic, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *CoRR*, abs/1811.12359, 2018.

[423] S. M. Loos, G. Irving, C. Szegedy, and C. Kaliszyk. Deep network guided proof search. In T. Eiter and D. Sands, editors, *Proc. 21st Int. Conf. on Logic for Programming, Artificial Intelligence and Reasoning (LPAR'17), Maun, Botswana, May 7-12*, volume 46 of *EPiC Series in Computing*, pages 85–105. EasyChair, 2017.

[424] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Scholkopf, and L. Bottou. Discovering causal signals in images. In *Proc. of the IEEE conf. on Computer Vision and Pattern Recognition*, pages 6979–6987, 2017.

[425] K. Loquin and O. Strauss. On the granularity of summative kernels. *Fuzzy Sets and Systems*, 159(15):1952–1972, 2008.

[426] S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conf. on Neural Information Processing Systems 2017, Dec. 4-9, Long Beach*, pages 4765–4774, 2017.

[427] C. Luo, S. Cai, W. Wu, Z. Jie, and K. Su. CCLS: an efficient local search algorithm for weighted maximum satisfiability. *IEEE Transactions on Computers*, 64(7):1830–1843, 2015.

[428] A. Mahinpei, J. Clark, I. Lage, F. Doshi-Velez, and W. Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

[429] R. Majumdar and V. Kuncak, editors. *Proc. 29th Int. Conf. on Computer Aided Verification (CAV'17), Heidelberg, July 24-28, Part I*, volume 10426 of *LNCS*. Springer, 2017.

[430] D. Malioutov and K. S. Meel. MLIC: A maxsat-based framework for learning interpretable classification rules. In Hooker [318], pages 312–327.

[431] S. Mallat. *Sciences des Données et Apprentissage en Grande Dimension*. Leçons Inaugurales du Collège de France. Fayard, Paris, 2018.

[432] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics.

[433] E. H. Mamdani and S. Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *Int. J. of Man-Machine Studies*, 7(1):1 – 13, 1975.

[434] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, and L. De Raedt. DeepProbLog: Neural probabilistic logic programming. In Bengio et al. [60], pages 3753–3763.

[435] R. Manhaeve, G. Marra, T. Demeester, S. Dumancic, A. Kimmig, and L. De Raedt. Neuro-symbolic AI = neural + logical + probabilistic AI. In P. Hitzler and M. K. Sarker, editors, *Neuro-Symbolic Artificial Intelligence: The State of the Art*, volume 342, pages 173–191. IOS Press, 2021.

[436] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *Proc. 7th Int. Conf. on Learning Representations (ICLR'19), New Orleans, May 6-9*. OpenReview.net, 2019.

[437] J. Marques-Silva. Logic-based explainability in machine learning. In L. E. Bertossi and G. Xiao, editors, *Reasoning Web. Causality, Explanations and Declarative Knowledge - 18th Int. Summer School 2022, Berlin, Germany, September 27-30, 2022, Tutorial Lectures*, volume 13759 of *LNCS*, pages 24–104. Springer, 2022.

[438] J. Marques-Silva, T. Gerspacher, M. C. Cooper, A. Ignatiev, and N. Narodytska. Explanations for monotonic classifiers. In M. Meila and T. Zhang, editors, *Proc. 38th Int. Conf. Mach. Learn. (ICML'21)*, PMLR, 139, 7469-7479, 2021.

[439] J. Marques-Silva and A. Ignatiev. Delivering trustworthy AI through formal XAI. In *Proc. 36th AAAI Conf. on Artificial Intelligence, AAAI 2022, 34th Conf. on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symp. on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, Feb. 22 - March 1*, pages 12342–12350. AAAI Press, 2022.

[440] J. Marques-Silva and A. Ignatiev. No silver bullet: interpretable ML models must be explained. *Frontiers Artif. Intell.*, 6, 2023.

[441] J. Marques-Silva, I. Lynce, and S. Malik. Conflict-driven clause learning SAT solvers. In A. Biere, M. Heule, H. van Maaren, and T. Walsh, editors, *Handbook of Satisfiability - Second Edition*, volume 336 of *Frontiers in Artificial Intelligence and Applications*, pages 133–182. IOS Press, 2021.

[442] P. Marquis. Compile! In B. Bonet and S. Koenig, editors, *Proc. 29th AAAI Conf. on Artificial Intelligence (AAAI'15)*, pages 4112–4118. AAAI Press, 2015.

[443] G. Marra, S. Dumancic, R. Manhaeve, and L. D. Raedt. From statistical relational to neural symbolic artificial intelligence: a survey. *CoRR*, abs/2108.11451, 2021. Revised version in Artif. Intell. 328: 104062 (2024).

[444] G. Marra and O. Kuzelka. Neural Markov logic networks. In C. P. de Campos, M. H. Maathuis, and E. Quaeghebeur, editors, *Proc. 37th Conf. on Uncertainty in Artificial Intelligence (UAI'21), Virtual Event, 27-30 July*, volume 161 of *Proc. of Machine Learning Research*, pages 908–917. AUAI Press, 2021.

[445] C. Marsala and B. Bouchon-Meunier. Quality of measures for attribute selection in fuzzy decision trees. In *Proc IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE'10), Barcelona, July 18-23*, pages 1–8, 2010.

[446] R. Martins, V. M. Manquinho, and I. Lynce. Open-wbo: A modular maxsat solver. In C. Sinz and U. Egly, editors, *Proc. 17th Int. Conf. on Theory and Applications of Satisfiability Testing (SAT'14)*, volume 8561 of *LNCS*, pages 438–445. Springer, 2014.

[447] P. J. Matos, J. Planes, F. Letombe, and J. Marques-Silva. A MAX-SAT algorithm portfolio. In M. Ghallab, C. D. Spyropoulos, N. Fakotakis, and N. M. Avouris, editors, *ECAI 2008 - 18th European Conference on Artificial Intelligence, Patras, Greece, July 21-25, 2008, Proceedings*, volume 178 of *Frontiers in Artificial Intelligence and Applications*, pages 911–912. IOS Press, 2008.

[448] G. Mauris. A review of relationships between possibility and probability representations of uncertainty in measurement. *IEEE Trans. Instrum. Meas.*, 62(3):622–632, 2013.

[449] C. Meilicke, M. Wudage Chekol, M. Fink, and H. Stuckenschmidt. Reinforced anytime bottom up rule learning for knowledge graph completion. *CoRR*, abs/2004.04412, 2020.

[450] N. Messai, M. Devignes, A. Napoli, and M. Smaïl-Tabbone. Many-valued concept lattices for conceptual clustering and information retrieval. In M. Ghallab, C. D. Spyropoulos, N. Fakotakis, and N. M. Avouris, editors, *Proc. 18th Europ. Conf. on Artificial Intelligence (ECAI'08), Patras, July 21-25*, volume 178 of *Frontiers in Artificial Intelligence and Applications*, pages 127–131. IOS Press, 2008.

[451] L. Miclet, S. Bayoudh, and A. Delhay. Analogical dissimilarity: definition, algorithms and two experiments in machine learning. *JAIR, 32*, pages 793–824, 2008.

[452] L. Miclet and H. Prade. Handling analogical proportions in classical logic and fuzzy logics settings. In *Proc. 10th Eur. Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'09)*, pages 638–650. Springer, LNCS 5590, 2009.

[453] T. Miller. “But why?" understanding explainable artificial intelligence. *ACM Crossroads*, 25(3):20–25, 2019.

[454] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[455] T. Miller. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, 36, 2021.

[456] P. Minervini, M. Bosnjak, T. Rocktäschel, and S. Riedel. Towards neural theorem proving at scale. *CoRR*, abs/1807.08204, 2018.

[457] S. Minton. Quantitative results concerning the utility of explanation-based learning. *Artificial Intelligence*, 42(2-3):363–391, 1990.

[458] S. Minton and J. G. Carbonell. Strategies for learning search control rules: An explanation-based approach. In *IJCAI*, pages 228–235, 1987.

[459] B. Mirkin. *Core Concepts in Data Analysis: Summarization, Correlation, Visualization*. Springer, 2011.

[460] M. Mirman, T. Gehr, and M. T. Vechev. Differentiable abstract interpretation for provably robust neural networks. In J. G. Dy and A. Krause, editors, *Proc. of the 35th Int. Conf. on Machine Learning, ICML'18), Stockholm, July 10-15*, volume 80 of *Proceedings of Machine Learning Research*, pages 3575–3583. PMLR, 2018.

[461] T. Mitchell. *Version spaces: An approach to concept learning.* PhD thesis, Stanford University, 1979.

[462] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[463] T. M. Mitchell. Version spaces: A candidate elimination approach to rule learning. In R. Reddy, editor, *Proc. 5th Int. Joint Conf. on Artificial Intelligence. Cambridge, MA, Aug. 22-25, 1977*, pages 305–310. William Kaufmann, 1977.

[464] T. M. Mitchell, R. M. Keller, and S. T. Kedar-Cabelli. Explanation-based generalization: A unifying view. *Machine learning*, 1:47–80, 1986.

[465] B. D. Mittelstadt, C. Russell, and S. Wachter. Explaining explanations in AI. In *FAT*, pages 279–288, 2019.

[466] V. Molek and I. Perfilieva. Scale-space theory, F-transform kernels and CNN realization. In *Advances in Computational Intelligence - Proc. 15th Int. Work-Conf. on Artificial Neural Networks, IWANN 2019, Gran Canaria, June 12-14, Part II*, pages 38–48, 2019.

[467] C. Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[468] D. R. Montalván Hernández, T. Centen, T. Krak, E. Quaeghebeur, and C. de Campos. Beyond tree-shaped credal probabilistic circuits. *Int. J. of Approximate Reasoning*, this issue:109047, 2023.

[469] G. Montavon, W. Samek, and K. Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[470] M. Mueller and S. Kramer. Integer linear programming models for constrained clustering. In *Discovery Science - Proc. 13th Int. Conf. DS'10, Canberra, Oct. 6-8*, pages 159–173, 2010.

[471] S. T. Mueller, R. R. Hoffman, W. J. Clancey, A. Emrey, and G. Klein. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *CoRR*, abs/1902.01876, 2019.

[472] S. Muggleton. Inverse entailment and Progol. *New Generation Comput.*, 13(3-4):245–286, 1995.

[473] S. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. *J. Log. Program.*, 19/20:629–679, 1994.

[474] S. H. Muggleton, U. Schmid, C. Zeller, A. Tamaddoni-Nezhad, and T. R. Besold. Ultra-strong machine learning: comprehensibility of programs learned with ILP. *Machine Learning*, 107(7):1119–1140, 2018.

[475] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In J. Fürnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 807–814. Omnipress, 2010.

[476] N. Narodytska. Formal analysis of deep binarized neural networks. In Lang [392], pages 5692–5696.

[477] N. Narodytska, A. Ignatiev, F. Pereira, and J. Marques-Silva. Learning optimal decision trees with SAT. In Lang [392], pages 1362–1368.

[478] N. Narodytska, S. P. Kasiviswanathan, L. Ryzhyk, M. Sagiv, and T. Walsh. Verifying properties of binarized deep neural networks. In *Proc. 32nd AAAI Conf. on Artificial Intelligence, (AAAI-18), New Orleans, Feb. 2-7*, pages 6615–6624, 2018.

[479] N. Narodytska, A. A. Shrotri, K. S. Meel, A. Ignatiev, and J. Marques-Silva. Assessing heuristic machine learning explanations with model counting. In Janota and Lynce [354], pages 267–278.

[480] N. Nghiem, C. Vrain, and T. Dao. Knowledge integration in deep clustering. In M. Amini, S. Canu, A. Fischer, T. Guns, P. K. Novak, and G. Tsoumakas, editors, *Proc. European conf. on Machine Learning and Knowledge Discovery in Databases (ECML PKDD'22), Grenoble, Sept. 19-23, Part I*, volume 13713 of *LNCS*, pages 174–190. Springer, 2022.

[481] H. T. Nguyen. On random sets and belief functions. *J. of Mathematical Analysis and Applications*, 65:531–542, 1978.

[482] V.-L. Nguyen and E. Hüllermeier. Multilabel classification with partial abstention: Bayes-optimal prediction under label independence. *J. of Artificial Intelligence Research*, 72:613–665, 2021.

[483] S. Nijssen. Bayes optimal classification for decision trees. In W. W. Cohen, A. McCallum, and S. T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 696–703. ACM, 2008.

[484] S. Nijssen and É. Fromont. Mining optimal decision trees from itemset lattices. In P. Berkhin, R. Caruana, and X. Wu, editors, *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, pages 530–539. ACM, 2007.

[485] S. Nijssen and É. Fromont. Optimal constraint-based decision tree induction from itemset lattices. *Data Min. Knowl. Discov.*, 21(1):9–51, 2010.

[486] J. Nin, A. Laurent, and P. Poncelet. Speed up gradual rule mining from stream data! A B-tree and owa-based approach. *J. Intell. Inf. Syst.*, 35(3):447–463, 2010.

[487] A. Ouali, A. Zimmermann, S. Loudni, Y. Lebbah, B. Crémilleux, P. Boizumault, and L. Loukil. Integer linear programming for pattern set mining; with an application to tiling. In *Proc. 21st Pacific-Asia conf. in Knowledge Discovery and Data Mining (PAKDD'17), Jeju, South Korea, May 23-26, Part II*, pages 286–299, 2017.

[488] A. Paliwal, S. M. Loos, M. N. Rabe, K. Bansal, and C. Szegedy. Graph representations for higher-order logic and theorem proving. *CoRR*, abs/1905.10006, 2019.

[489] R. B. Palm, U. Paquet, and O. Winther. Recurrent relational networks. In Bengio et al. [60], pages 3372–3382.

[490] P. Panda and K. Roy. Explainable learning: Implicit generative modelling during training for adversarial robustness. *CoRR*, abs/1807.02188, 2018.

[491] E. Parisotto, A. Mohamed, R. Singh, L. Li, D. Zhou, and P. Kohli. Neuro-symbolic program synthesis. In *Proc. 5th Int. Conf. on Learning Representations (ICLR'17), Toulon, Apr. 24-26* [339].

[492] S. Parsons. *Qualitative Approaches for Reasoning Under Uncertainty*. MIT Press, 2001.

[493] Z. Pawlak. *Rough Sets. Theoretical Aspects of. Reasoning about Data*. Kluwer Acad. Publ., Dordrecht, 1991.

[494] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[495] J. Pearl. *Causality*. Cambridge university press, 2009.

[496] J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[497] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[498] W. Pedrycz. Conditional fuzzy clustering in the design of radial basis function neural networks. *IEEE Trans. Neural Networks*, 9(4):601–612, 1998.

[499] G. Pinkas. Symmetric neural networks and propositional logic satisfiability. *Neural Computation*, 3(2):282–291, 1991.

[500] G. Pinkas. Reasoning, nonmonotonicity and learning in connectionist networks that capture propositional knowledge. *Artificial Intelligence*, 77(2):203 – 247, 1995.

[501] G. Pinkas and S. Cohen. High-order networks that learn to satisfy logic constraints. *J. of Applied Logics - IfCoLog J. of Logics and their Applications*, 6(4):653–694, 2019.

[502] G. Plotkin. A note on inductive generalization. In *Machine Intelligence*, volume 5, pages 153–163. Edinburgh University Press, 1970.

[503] H. Prade and G. Richard. From analogical proportion to logical proportions. *Logica Universalis*, 7(4):441–505, 2013.

[504] H. Prade and G. Richard. Analogical proportions: From equality to inequality. *Int. J. Approx. Reasoning*, 101:234–254, 2018.

[505] H. Prade, A. Rico, and M. Serrurier. Elicitation of Sugeno integrals: A version space learning perspective. In J. Rauch, Z. W. Ras, P. Berka, and T. Elomaa, editors, *Proc. 18th Int. Symp. on Foundations of Intelligent Systems (ISMIS'09), Prague, Sept. 14-17*, volume 5722 of *LNCS*, pages 392–401. Springer, 2009.

[506] H. Prade, A. Rico, M. Serrurier, and E. Raufaste. Elicitating Sugeno integrals:Methodology and a case study. In C. Sossai and G. Chemello, editors, *Proc. 10th Eur. Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'09), Verona, July 1-3*, volume 5590 of *LNCS*, pages 712–723. Springer, 2009.

[507] M. O. R. Prates, P. H. C. Avelar, H. Lemos, L. C. Lamb, and M. Y. Vardi. Learning to solve NP-complete problems: A graph neural network for decision TSP. In *Proc. 33rd AAAI Conf. on Artificial Intelligence (AAAI'19), The 31st Innovative Applications of Artificial Intelligence Conf., IAAI 2019, The 9th AAAI Symp. on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Jan. 27 - Feb. 1* [1], pages 4731–4738.

[508] S. Prestwich and N. Wilson. A statistical approach to learning constraints. *Int. J. of Approximate Reasoning*, this issue:109184, 2024.

[509] A. Procopio, G. Cesarelli, L. Donisi, A. Merola, F. Amato, and C. Cosentino. Combined mechanistic modeling and machine-learning approaches in systems biology–a systematic literature review. *Computer methods and programs in biomedicine*, page 107681, 2023.

[510] C. Pryor, C. Dickens, E. Augustine, A. Albalak, W. Wang, and L. Getoor. Neupsl: Neural probabilistic soft logic, 2023.

[511] L. Pulina and A. Tacchella. An abstraction-refinement approach to verification of artificial neural networks. In T. Touili, B. Cook, and P. B. Jackson, editors, *Proc. 22nd Int. Conf. on Computer Aided Verification (CAV'10), Edinburgh, July 15-19*, volume 6174 of *LNCS*, pages 243–257. Springer, 2010.

[512] C. Qin, K. D. Dvijotham, B. O'Donoghue, R. Bunel, R. Stanforth, S. Gowal, J. Uesato, G. Swirszcz, and P. Kohli. Verification of non-linear specifications for neural networks. In *Proc. 7th Int. Conf. on Learning Representations (ICLR'19), New Orleans, May 6-9*. OpenReview.net, 2019.

[513] J. R. Quinlan. Learning first-order definitions of functions. *CoRR*, cs.AI/9610102, 1996.

[514] B. Quost, T. Denœux, and S. Li. Parametric classification with soft labels using the evidential EM algorithm: linear discriminant analysis versus logistic regression. *Advances in Data Analysis and Classification*, 11(4):659–690, Dec 2017.

[515] B. Quost, M.-H. Masson, and T. Denœux. Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules. *Int. J. of Approximate Reasoning*, 52(3):353–374, 2011.

[516] L. D. Raedt, A. Dries, T. Guns, and C. Bessiere. Learning constraint satisfaction problems: An ILP perspective. In C. Bessiere, L. D. Raedt, L. Kotthoff, S. Nijssen, B. O'Sullivan, and D. Pedreschi, editors, *Data Mining and Constraint Programming - Foundations of a Cross-Disciplinary Approach*, volume 10101 of *LNCS*, pages 96–112. Springer, 2016.

[517] L. D. Raedt, A. Passerini, and S. Teso. Learning constraints from examples. In S. A. McIlraith and K. Q. Weinberger, editors, *Proc. 32nd AAAI Conf. on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symp. on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Feb. 2-7*, pages 7965–7970. AAAI Press, 2018.

[518] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. of Computational physics*, 378:686–707, 2019.

[519] V. V. Ramasesh, E. Dyer, and M. Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *ICLR-2021*, 2021.

[520] E. Ramasso and T. Denœux. Making use of partial knowledge about hidden states in HMMs: an approach based on belief functions. *IEEE Transactions on Fuzzy Systems*, 21(6):1–11, 2013.

[521] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Krishnapuram et al. [381], pages 1135–1144.

[522] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In S. A. McIlraith and K. Q. Weinberger, editors, *Proc. 32nd AAAI Conf. on Artificial Intelligence (AAAI-18), New Orleans, Feb. 2-7*, pages 1527–1535, 2018.

[523] M. Richardson and P. M. Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.

[524] S. Riedel, L. Yao, and A. McCallum. Modeling relations and their mentions without labeled text. In *Joint European conf. on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer, 2010.

[525] S. Riedel, L. Yao, A. McCallum, and B. M. Marlin. Relation extraction with matrix factorization and universal schemas. In L. Vanderwende, H. D. III, and K. Kirchhoff, editors, *Proc. Human Language Technologies: Conf. e North American Chapter of the Association of Computational Linguistics (HLT-NAACL'13), June 9-14, Atlanta*, pages 74–84, 2013.

[526] T. Rocktäschel and S. Riedel. Learning knowledge base inference with neural theorem provers. In J. Pujara, T. Rocktäschel, D. Chen, and S. Singh, editors, *Proc. 5th Workshop on Automated Knowledge Base Construction, AKBC@NAACL-HLT 2016, San Diego, CA, USA, June 17, 2016*, pages 45–50. The Association for Computer Linguistics, 2016.

[527] T. Rocktäschel and S. Riedel. End-to-end differentiable proving. In Guyon et al. [294], pages 3788–3800.

[528] C. Rodriguez, V. M. Bordini, S. Destercke, and B. Quost. Self learning using venn-abers predictors. In *Conformal and Probabilistic Prediction with Applications*, pages 234–250. PMLR, 2023.

[529] P. Rodriguez, M. Caccia, A. Lacoste, L. Zamparo, I. Laradji, L. Charlin, and D. Vazquez. Beyond trivial counterfactual explanations with diverse valuable explanations. In *Int. Conf in Computer Vision (ICCV)*, 2021.

[530] G. Rogova. Combining the results of several neural network classifiers. *Neural Networks*, 7(5):777–781, 1994.

[531] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[532] A. S. Ross and F. Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proc. 32nd AAAI Conf. on Artificial Intelligence, (AAAI-18), New Orleans, Feb. 2-7*, pages 1660–1669, 2018.

[533] F. Rossi, P. van Beek, and T. Walsh, editors. *Handbook of Constraint Programming*, volume 2 of *Foundations of Artificial Intelligence*. Elsevier, 2006.

[534] M. Rousset and B. Safar. Negative and positive explanations in expert. *Applied Artificial Intelligence*, 1(1):25–38, 1987.

[535] W. Ruan, X. Huang, and M. Kwiatkowska. Reachability analysis of deep neural networks with provable guarantees. In Lang [392], pages 2651–2659.

[536] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[537] D. E. Rumelhart, J. L. McClelland, P. R. Group, et al. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press, 1986.

[538] A. Ruschel, A. Colombini Gusmão, and F. Gagliardi Cozman. Explaining answers generated by knowledge graph embeddings. *Int. J. of Approximate Reasoning*, this issue:109183, 2024.

[539] B. Russell. *The Problems of Philosophy. Chap. VI. On induction*. Home Univ. Libr.; Oxford Univ. Pr., 1959, 1912.

[540] S. J. Russell. Unifying logic and probability. *Commun. ACM*, 58(7):88–97, 2015.

[541] H. X. S. H. Huanga. Extract intelligible and concise fuzzy rules from neural networks. *Fuzzy Sets and Systems*, 132:233–243, 2002.

[542] W. C. Salmon. *Causality and Explanation*. Oxford University Press, 1998.

[543] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K. Müller, editors. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *LNCS*. Springer, 2019.

[544] W. Samek and K. Müller. Towards explainable artificial intelligence. In Samek et al. [543], pages 5–22.

[545] E. Sanchez. Resolution of composite fuzzy relation equations. *Information and control*, 30(1):38–48, 1976.

[546] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi. ATOMIC: An atlas of machine commonsense for if-then reasoning. In *Proc. of the 33rd AAAI conf. on artificial intelligence, Honolulu, Jan. 27 - Feb.1*, pages 3027–3035, 2019.

[547] D. Saxton, E. Grefenstette, F. Hill, and P. Kohli. Analysing mathematical reasoning abilities of neural models. In *Proc. 7th Int. Conf. on Learning Representations (ICLR'19), New Orleans, May 6-9*. OpenReview.net, 2019.

[548] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Trans. Neural Networks*, 20(1):61–80, 2009.

[549] T. Schiex, H. Fargier, and G. Verfaillie. Valued constraint satisfaction problems: Hard and easy problems. In *Proc. 14th Int. Joint Conf. on Artificial Intelligence (IJCAI'95)*, pages 631–639. Morgan Kaufmann, 1995.

[550] S. Schockaert. Embeddings as epistemic states: Limitations on the use of pooling operators for accumulating knowledge. *Int. J. of Approximate Reasoning*, this issue:108981, 2023.

[551] S. Schockaert and H. Prade. Interpolative and extrapolative reasoning in propositional theories using qualitative knowledge about conceptual spaces. *Artif. Intell.*, 202:86–131, 2013.

[552] J. Schumann and S. D. Nelson. Toward v&v of neural network based controllers. In D. Garlan, J. Kramer, and A. L. Wolf, editors, *Proceedings of the First Workshop on Self-Healing Systems, WOSS 2002, Charleston, South Carolina, USA, November 18-19, 2002*, pages 67–72. ACM, 2002.

[553] N. Schwind, K. Inoue, and P. Marquis. Editing boolean classifiers: A belief change perspective. In B. Williams, Y. Chen, and J. Neville, editors, *Proc. 37th AAAI conf. on Artificial Intelligence, (AAAI'23), Thirty-Fifth conf. on Innovative Applications of Artificial Intelligence, IAAI 2023, 13th Symp. on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 6516–6524. AAAI Press, 2023.

[554] D. Selsam and N. Bjørner. Guiding high-performance SAT solvers with unsat-core predictions. In Janota and Lynce [354], pages 336–353.

[555] D. Selsam, M. Lamm, B. Bünz, P. Liang, L. de Moura, and D. L. Dill. Learning a SAT solver from single-bit supervision. *CoRR*, abs/1802.03685, 2018.

[556] D. Selsam, M. Lamm, B. Bünz, P. Liang, L. de Moura, and D. L. Dill. Learning a SAT solver from single-bit supervision. In *Proc. 7th Int. Conf. on Learning Representations (ICLR'19), New Orleans, May 6-9*. OpenReview.net, 2019.

[557] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Inter. J. of Computer Vision*, 128(2):336–359, oct 2019.

[558] L. Serafini and A. S. d'Avila Garcez. Learning and reasoning with logic tensor networks. In G. Adorni, S. Cagnoni, M. Gori, and M. Maratea, editors, *AI*IA 2016: Advances in Artificial Intelligence - Proc. XVth Int. Conf. of the Italian Association for Artificial Intelligence, Genova, Nov. 29 - Dec. 1*, volume 10037 of *LNCS*, pages 334–348. Springer, 2016.

[559] M. Serrurier, D. Dubois, H. Prade, and T. Sudkamp. Learning fuzzy rules with their implication operators. *Data Knowl. Eng.*, 60(1):71–89, 2007.

[560] M. Serrurier, F. Mamalet, T. Fel, L. Béthune, and T. Boissin. On the explainable properties of 1-lipschitz neural networks: An optimal transport perspective. In *Advances in Neural Information Processing Systems (NeurIPS) 2023*, 2023.

[561] M. Serrurier and H. Prade. Introducing possibilistic logic in ILP for dealing with exceptions. *Artif. Intell.*, 171(16-17):939–950, 2007.

[562] M. Serrurier and H. Prade. An informational distance for estimating the faithfulness of a possibility distribution, viewed as a family of probability distributions, with respect to data. *Int. J. Approx. Reasoning*, 54(7):919–933, 2013.

[563] M. Serrurier and H. Prade. Entropy evaluation based on confidence intervals of frequency estimates : Application to the learning of decision trees. In F. R. Bach and D. M. Blei, editors, *Proc. 32nd Int. Conf. on Machine Learning (ICML'15), Lille, July 6-11*, volume 37 of *JMLR Workshop and conf. proc.*, pages 1576–1584. JMLR.org, 2015.

[564] S. A. Seshia, A. Desai, T. Dreossi, D. J. Fremont, S. Ghosh, E. Kim, S. Shivakumar, M. Vazquez-Chanlatte, and X. Yue. Formal specification for deep neural networks. In Lahiri and Wang [389], pages 20–34.

[565] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J., 1976.

[566] G. Shafer and V. Vovk. A tutorial on conformal prediction. *J. of Machine Learning Research*, 9(Mar):371–421, 2008.

[567] E. Shapiro. Inductive inference of theories from facts, 1981. Tech. report 192, Depart. Computer Sci., Yale Univ. Reprinted in Lassez, J.-L.; Plotkin, G., eds.: Computational Logic : Essays in Honor of Alan Robinson, MIT Press, pp. 199–254, 1991.

[568] J. W. Shavlik and T. G. Dietterich. *Readings in machine learning*. Morgan Kaufmann, 1990.

[569] J. Shen, Y. Qu, W. Zhang, and Y. Yu. Wasserstein distance guided representation learning for domain adaptation. *arXiv preprint arXiv:1707.01217*, 2017.

[570] P. P. Shenoy. Conditional independence in valuation-based systems. *Int. J. Approx. Reasoning*, 10(3):203–234, 1994.

[571] A. Shih, A. Choi, and A. Darwiche. A symbolic approach to explaining bayesian network classifiers. In Lang [392], pages 5103–5111.

[572] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Y. Bengio and Y. LeCun, editors, *Proc. 2nd Int. Conf. on Learning Representations, ICLR 2014, Banff, Apr. 14-16, Workshop Track*, 2014.

[573] G. Singh, T. Gehr, M. Mirman, M. Püschel, and M. T. Vechev. Fast and effective robustness certification. In Bengio et al. [60], pages 10825–10836.

[574] G. Singh, T. Gehr, M. Püschel, and M. T. Vechev. An abstract domain for certifying neural networks. *PACMPL*, 3(POPL):41:1–41:30, 2019.

[575] G. Singh, T. Gehr, M. Püschel, and M. T. Vechev. Boosting robustness certification of neural networks. In *Proc. 7th Int. Conf. on Learning Representations (ICLR'19), New Orleans, May 6-9*. OpenReview.net, 2019.

[576] R. Singh, J. P. Near, V. Ganesh, and M. Rinard. AvatarSAT: An auto-tuning boolean SAT solver. Technical Report MIT-CSAIL-TR-2009-039, MIT, 2009.

[577] L. Sixt, M. Granz, and T. Landgraf. When explanations lie: Why many modified bp attributions fail. In *Proc 37th Int. Conf. on Machine Learning*, 2020.

[578] A. Skowron and H. S. Nguyen. Boolean reasoning scheme with some applications in data mining. In J. M. Zytkow and J. Rauch, editors, *Proc. 3rd Europ. Conf. on Principles of Data Mining and Knowledge Discovery, (PKDD'99), Prague, Sept. 15-18*, volume 1704 of *LNCS*, pages 107–115. Springer, 1999.

[579] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proc. of the AAAI/ACM conf. on AI, Ethics, and Society*, pages 180–186, 2020.

[580] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise, 2017.

[581] R. Socher, D. Chen, C. D. Manning, and A. Y. Ng. Reasoning with neural tensor networks for knowledge base completion. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, *Proc. of the 27th Annual conf. on Neural Information Processing Systems (NIPS 2013)*, pages 926–934, 2013.

[582] G. Sourek, V. Aschenbrenner, F. Zelezný, S. Schockaert, and O. Kuzelka. Lifted relational neural networks: Efficient learning of latent relational structures. *J. Artif. Intell. Res.*, 62:69–100, 2018.

[583] J. F. Sowa. *Conceptual Structures : Information Processing in Mind and Machine*. Addison-Wesley, 1984.

[584] R. Srinivasan and A. Chander. Biases in ai systems. *Communications of the ACM*, 64(8):44–49, 2021.

[585] S. Srinivasan, C. Dickens, E. Augustine, G. Farnadi, and L. Getoor. A taxonomy of weight learning methods for statistical relational learning. *Mach. Learn.*, 111(8):2799–2838, 2022.

[586] O. Strauss and A. Rico. Macsum aggregation learning and missing values. In Z. Bouraoui and S. Vesic, editors, *Proc. 17th Europ. Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'23), Arras, Sept. 19-22, 2023*, volume 14294 of *LNCS*, pages 453–463. Springer, 2024.

[587] O. Strauss, A. Rico, and Y. Hmidy. Macsum: a new interval-valued linear operator. *Int. J. of Approx. Reas.*, 145:121–138, 2022.

[588] N. Stroppa and F. Yvon. Analogical learning and formal proportions: Definitions and methodological issues. Technical Report D004, ENST-Paris, 2005.

[589] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks, 2017.

[590] M. Svatos, S. Schockaert, J. Davis, and O. Kuzelka. Strike: Rule-driven relational learning using stratified k-entailment. In G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, and J. Lang, editors, *Proc. 24th Europ. Conf. on Artificial Intelligence (ECAI'20), Santiago de Compostela, 29 Aug.-8 Sept.*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 1515–1522. IOS Press, 2020.

[591] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In Y. Bengio and Y. LeCun, editors, *Proc. 2nd Int. Conf. on Learning Representations (ICLR'14), Banff, April 14-16*, 2014.

[592] T. Takagi and M. Sugeno. Fuzzy identification of systems and its application to modelling and control. *IEEE Trans. Systems, Man, and Cybernetics*, 15(1):11)–132, 1985.

[593] G. Tao, S. Ma, Y. Liu, and X. Zhang. Attacks meet interpretability: Attribute-steered detection of adversarial samples. In Bengio et al. [60], pages 7728–7739.

[594] P. R. Thagard. The best explanation: Criteria for theory choice. *The J. of Philosophy*, 75(2):76–92, 1978.

[595] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[596] I. Tiddi and S. Schlobach. Knowledge graphs as tools for explainable machine learning: A survey. *Artif. Intell.*, 302:103627, 2022.

[597] R. Tomsett, A. Widdicombe, T. Xing, S. Chakraborty, S. Julier, P. Gurram, R. M. Rao, and M. B. Srivastava. Why the failure? how adversarial examples can provide insights for interpretable machine learning. In *Proc. 21st Int. Conf. on Information Fusion (FUSION'18), Cambridge, UK, July 10-13*, pages 838–845. IEEE, 2018.

[598] Z. Tong, P. Xu, and T. Denœux. An evidential classifier based on Dempster-Shafer theory and deep learning. *Neurocomputing*, 450:275–293, 2021.

[599] Z. Tong, P. Xu, and T. Denœux. Evidential fully convolutional network for semantic segmentation. *Applied Intelligence*, 51:6376–6399, 2021.

[600] K. Toutanova, D. Chen, P. Pantel, H. Poon, P. Choudhury, and M. Gamon. Representing text for joint embedding of text and knowledge bases. In *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15), Lisbon, Sept. 17-21*, pages 1499–1509, 2015.

[601] G. G. Towell and J. W. Shavlik. Extracting refined rules from knowledge-based neural networks. *Mach. Learn.*, 13:71–101, 1993.

[602] G. G. Towell and J. W. Shavlik. Knowledge-based artificial neural networks. *Artif. Intell.*, 70(1-2):119–165, 1994.

[603] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard. Complex embeddings for simple link prediction. In *Int. Conf. on Machine Learning*, pages 2071–2080, 2016.

[604] D. C. Tsouros, S. Berden, and T. Guns. Guided bottom-up interactive constraint acquisition. In R. H. C. Yap, editor, *Proc. 29th Int. Conf. on Principles and Practice of Constraint Programming (CP'23), Aug. 27-31, Toronto*, volume 280 of *LIPIcs*, pages 36:1–36:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023.

[605] D. C. Tsouros, S. Berden, and T. Guns. Learning to learn in interactive constraint acquisition. In M. J. Wooldridge, J. G. Dy, and S. Natarajan, editors, *Proc. 38th AAAI Conf. on Artificial Intelligence, (AAAI'24), 36th Conf. on Innovative Applications of Artificial Intelligence, IAAI 2024, 14th Symp. on Educational Advances in Artificial Intelligence, EAAI 2014, Feb. 20-27, Vancouver*, pages 8154–8162. AAAI Press, 2024.

[606] D. C. Tsouros, K. Stergiou, and C. Bessiere. Structure-driven multiple constraint acquisition. In T. Schiex and S. de Givry, editors, *Proc. 25th Int. Conf. on Principles and Practice of Constraint Programming (CP'19), Stamford, CT, Sept. 30 - Oct. 4*, volume 11802 of *LNCS*, pages 709–725. Springer, 2019.

[607] D. C. Tsouros, K. Stergiou, and C. Bessiere. Omissions in constraint acquisition. In H. Simonis, editor, *Proc. 26th Int. Conf. on Principles and Practice of Constraint Programming (CP'20), Louvain-la-Neuve, Sept. 7-11*, volume 12333 of *LNCS*, pages 935–951. Springer, 2020.

[608] D. C. Tsouros, K. Stergiou, and P. G. Sarigiannidis. Efficient methods for constraint acquisition. In J. N. Hooker, editor, *Proc. 24th Int. Conf. Principles and Practice of Constraint Programming (CP'18), Lille, Aug. 27-31*, volume 11008 of *LNCS*, pages 373–388. Springer, 2018.

[609] J. Urban, J. Vyskocil, and P. Stepánek. Malecop machine learning connection prover. In K. Brünnler and G. Metcalfe, editors, *Proc. 20th Int. Conf. on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX'11), Bern, July 4-8*, volume 6793 of *LNCS*, pages 263–277. Springer, 2011.

[610] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[611] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2013.

[612] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

[613] H. Venkateswara, S. Chakraborty, and S. Panchanathan. Deep-learning systems for domain adaptation in computer vision: Learning transferable feature representations. *IEEE Signal Processing Magazine*, 34(6):117–129, 2017.

[614] H. Verhaeghe, S. Nijssen, G. Pesant, C.-G. Quimper, and P. Schaus. Learning optimal decision trees using constraint programming. *Constraints*, 2019.

[615] S. Verma, J. P. Dickerson, and K. Hines. Counterfactual explanations for machine learning: A review. *CoRR*, abs/2010.10596, 2020.

[616] S. Verwer and Y. Zhang. Learning decision trees with flexible constraints and objectives using integer optimization. In D. Salvagnin and M. Lombardi, editors, *Proc. 14th Int. Conf. on Integration of AI and OR Techniques in Constraint Programming (CPAIOR'17), Padua, June 5-8*, volume 10335 of *LNCS*, pages 94–103. Springer, 2017.

[617] S. Verwer and Y. Zhang. Learning optimal classification trees using a binary linear program formulation. In *Proc. 33rd AAAI Conf. on Artificial Intelligence (AAAI'19), The 31st Innovative Applications of Artificial Intelligence Conf., IAAI 2019, The 9th AAAI Symp. on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Jan. 27 - Feb. 1* [1], pages 1625–1632.

[618] L. Vilnis and A. McCallum. Word representations via gaussian embedding. In *Proc. Int. Conf. on Learning Representations*, 2015.

[619] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conf. on Neural Information Processing Systems, Dec. 7-12, Montreal*, pages 2692–2700, 2015.

[620] G. H. von Wright. *Norm and Action*. Routledge and Keagan, 1963.

[621] S. Wachter, B. Mittelstadt, and R. Chris. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard J. of Law & Technology*, 2018.

[622] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[623] P. Walley and S. Moral. Upper probabilities based only on the likelihood function. *J. of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(4):831–847, 1999.

[624] C. Wang, R. Bunel, K. Dvijotham, P. Huang, E. Grefenstette, and P. Kohli. Knowing when to stop: Evaluation and verification of conformity to output-size specifications. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'19), Long Beach, June 16-20*, pages 12260–12269. Computer Vision Foundation / IEEE, 2019.

[625] H. Wang, F. Zhang, X. Xie, and M. Guo. DKN: deep knowledge-aware network for news recommendation. In P. Champin, F. Gandon, M. Lalmas, and P. G. Ipeirotis, editors, *Proc. of the 2018 World Wide Web conf. on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1835–1844. ACM, 2018.

[626] M. Wang, Y. Tang, J. Wang, and J. Deng. Premise selection for theorem proving by deep graph embedding. In Guyon et al. [294], pages 2786–2796.

[627] P. Wang, P. L. Donti, B. Wilder, and J. Z. Kolter. SATNet: bridging deep learning and logical reasoning using a differentiable satisfiability solver. In *Proc 36th Int. Conf. on Machine Learning (ICML'19), Long Beach June 9-15*, pages 6545–6554, 2019.

[628] P. Wang and N. Vasconcelos. Scout: Self-aware discriminant counterfactual explanations. In *The IEEE/CVF conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[629] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021.

[630] X. Wang, X. He, Y. Cao, M. Liu, and T. Chua. KGAT: knowledge graph attention network for recommendation. In A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, and G. Karypis, editors, *Proc. 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, Aug. 4-8*, pages 950–958. ACM, 2019.

[631] X. Wang, D. Wang, C. Xu, X. He, Y. Cao, and T. Chua. Explainable reasoning over knowledge graphs for recommendation. In *Proc. 33rd AAAI Conf. on Artificial Intelligence (AAAI'19), 31st Innovative Applications of Artificial Intelligence conf. (IAAI'19), 9th AAAI Symp. on Educational Advances in Artificial Intelligence (EAAI'19), Honolulu, Jan. 27 - Feb. 1*, pages 5329–5336. AAAI Press, 2019.

[632] S. Webb, T. Rainforth, Y. W. Teh, and M. P. Kumar. A statistical approach to assessing neural network robustness. In *Proc. 7th Int. Conf. on Learning Representations (ICLR'19), New Orleans, May 6-9*. OpenReview.net, 2019.

[633] A. Weller. Transparency: Motivations and challenges. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and M. Klaus-Robert, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *LNCS*, pages 23–40. Springer, 2019.

[634] W. Wen, J. Callahan, and M. Napolitano. Towards developing verifiable neural network controller. In *Workshop on AI for Aeronautics and Space*, 1996.

[635] P. West, C. Bhagavatula, J. Hessel, J. Hwang, L. Jiang, R. Le Bras, X. Lu, S. Welleck, and Y. Choi. Symbolic knowledge distillation: from general language models to commonsense models. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proc. of the 2022 conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States, July 2022. Association for Computational Linguistics.

[636] D. Whalen. Holophrasm: a neural automated theorem prover for higher-order logic. *CoRR*, abs/1608.02644, 2016.

[637] N. Wilson. An efficient upper approximation for conditional preference. In G. Brewka, S. Coradeschi, A. Perini, and P. Traverso, editors, *Proc. of the 17th European conf. on Artificial Intelligence (ECAI 2006)*, Frontiers in Artificial Intelligence and Applications. IOS Press, 2006.

[638] N. Wilson. Efficient inference for expressive comparative preference language. In C. Boutilier, editor, *Proc. 21st Int. Joint Conf. on Artificial Intelligence (IJCAI'09)*, pages 961–966, 2009.

[639] N. Wilson. Preference inference based on lexicographic models. In T. Schaub, G. Friedrich, and B. O'Sullivan, editors, *Proc. 21st Europ. Conf. on Artificial Intelligence (ECAI14)*, volume 263 of *Frontiers in Artificial Intelligence and Applications*, pages 921–926. IOS Press, 2014.

[640] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, et al. A survey on large language models for recommendation. *arXiv preprint arXiv:2305.19860*, 2023.

[641] M. Wu, H. Wu, and C. W. Barrett. Verix: Towards verified explainability of deep neural networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conf. on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, Dec. 10 - 16*, 2023.

[642] H. Xiao, M. Huang, L. Meng, and X. Zhu. SSP: Semantic space projection for knowledge graph embedding with text descriptions. In S. Singh and S. Markovitch, editors, *Proc. of the 31st AAAI Conf. on Artificial Intelligence, Feb. 4-9, San Francisco*, pages 3104–3110, 2017.

[643] R. Xie, Z. Liu, J. Jia, H. Luan, and M. Sun. Representation learning of knowledge graphs with entity descriptions. In *Proceedigs of AAAI*, pages 2659–2665, 2016.

[644] Y. Xie, Z. Xu, K. Meel, M. S. Kankanhalli, and H. Soh. Semantically-regularized logic graph embeddings. *CoRR*, abs/1909.01161, 2019.

[645] Y. Xie, Z. Xu, K. S. Meel, M. S. Kankanhalli, and H. Soh. Embedding symbolic knowledge into deep networks. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual conf. on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4235–4245, 2019.

[646] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *NLPCC*, pages 563–574, 2019.

[647] H. Xu, S. Koenig, and T. K. S. Kumar. Towards effective deep learning for constraint satisfaction problems. In Hooker [318], pages 588–597.

[648] J. Xu, Z. Zhang, T. Friedman, Y. Liang, and G. V. den Broeck. A semantic loss function for deep learning with symbolic knowledge. In *Proc. 35th Int. Conf. on Machine Learning*, pages 5498–5507, 2018.

[649] K. Xu, S. Liu, P. Zhao, P. Chen, H. Zhang, D. Erdogmus, Y. Wang, and X. Lin. Structured adversarial attack: Towards general implementation and better interpretability. *CoRR*, abs/1808.01664, 2018.

[650] L. Xu, F. Hutter, H. H. Hoos, and K. Leyton-Brown. SATzilla: Portfolio-based algorithm selection for SAT. *J. Artif. Intell. Res.*, 32:565–606, 2008.

[651] L. Xu, A. Krzyzak, and C. Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Trans. on Systems, Man and Cybernetics*, 22(3):418–435, 1992.

[652] P. Xu, F. Davoine, H. Zha, and T. Denœux. Evidential calibration of binary SVM classifiers. *Int. J. of Approximate Reasoning*, 72:55–70, 2016.

[653] B. B. Yaghlane and K. Mellouli. Inference in directed evidential networks based on the transferable belief model. *Int. J. Approx. Reasoning*, 48(2):399–418, 2008.

[654] F. Yang, Z. Yang, and W. W. Cohen. Differentiable learning of logical rules for knowledge base reasoning. In Guyon et al. [294], pages 2319–2328.

[655] K. Yang and J. Deng. Learning to prove theorems via interacting with proof assistants. In K. Chaudhuri and R. Salakhutdinov, editors, *Proc 36th Int. Conf. on Machine Learning (ICML'19), Long Beach June 9-15*, volume 97 of *Proc. of Machine Learning Research*, pages 6984–6994. PMLR, 2019.

[656] Z. Yang, F. Wang, Z. Chen, G. Wei, and T. Rompf. Graph neural reasoning for 2-quantified boolean formula solvers. *CoRR*, abs/1904.12084, 2019.

[657] Y. Yao. Three-way granular computing, rough sets, and formal concept analysis. *Int. J. Approx. Reason.*, 116:106–125, 2020.

[658] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proc. of the 2021 conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online, June 2021. Association for Computational Linguistics.

[659] K. Yoon, R. Liao, Y. Xiong, L. Zhang, E. Fetaya, R. Urtasun, R. S. Zemel, and X. Pitkow. Inference in probabilistic graphical models by graph neural networks. In *Proc. 6th Int. Conf. on Learning Representations (ICLR'18), Vancouver, Apr. 30 - May 3*, 2018.

[660] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[661] M. Yuksekgonul, M. Wang, and J. Zou. Post-hoc concept bottleneck models. In *The 11th Int. Conf. on Learning Representations*, 2022.

[662] L. A. Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Trans. Systems, Man, and Cybernetics*, 3(1):28–44, 1973.

[663] L. A. Zadeh. A theory of approximate reasoning. In J. E. Hayes, D. Mitchie, and L. L. Mikulich, editors, *Machine intelligence, Vol. 9*, pages 149–194. Ellis Horwood, 1979.

[664] L. A. Zadeh. The calculus of fuzzy if-then rules. *AI Expert*, 7(3):22–27, 1992.

[665] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proc. 26th Int. Conf. on World Wide Web*, WWW '17. Int. World Wide Web Conf. Steering Committee, Apr. 2017.

[666] M. Zaffalon, A. Antonucci, R. Cabañas, D. Huber, and D. Azzimonti. Efficient computation of counterfactual bounds. *Int. J. of Approximate Reasoning*, this issue:109111, 2024.

[667] R. R. Zakrzewski. Verification of a trained neural network accuracy. In *Int. Joint Conf. on Neural Networks (IJCNN)*, pages 1657–1662. IEEE, 2001.

[668] F. Zelezný and N. Lavrac. Propositionalization-based relational subgroup discovery with rsd. *Machine Learning*, 62(1-2):33–63, 2006.

[669] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In S. Dasgupta and D. McAllester, editors, *Proc. 30th Int. Conf. on Machine Learning*, volume 28 (3) of *Proc. of Machine Learning Research*, pages 325–333, Atlanta, 17–19 Jun 2013. PMLR.

[670] F. Zhang, N. J. Yuan, D. Lian, X. Xie, and W. Ma. Collaborative knowledge base embedding for recommender systems. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, and R. Rastogi, editors, *Pro. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Francisco, Aug. 13-17*, pages 353–362. ACM, 2016.

[671] H. Zhang, T. Zhan, S. Basu, and I. Davidson. A framework for deep constrained clustering. *Data Min. Knowl. Discov.*, 35(2):593–620, 2021.

[672] L. Zhang, G. Rosenblatt, E. Fetaya, R. Liao, W. E. Byrd, M. Might, R. Urtasun, and R. S. Zemel. Neural guided constraint logic programming for program synthesis. In Bengio et al. [60], pages 1744–1753.

[673] L. Zhang, G. Rosenblatt, E. Fetaya, R. Liao, W. E. Byrd, R. Urtasun, and R. S. Zemel. Leveraging constraint logic programming for neural guided program synthesis. In *Proc. 6th Int. Conf. on Learning Representations (ICLR'18), Vancouver, Apr. 30 - May 3*. OpenReview.net, 2018.

[674] X. Zhang, A. Bosselut, M. Yasunaga, H. Ren, P. Liang, C. D. Manning, and J. Leskovec. Greaselm: Graph reasoning enhanced language models. In *The 10th Int. Conf. on Learning Representations, ICLR 2022, Virtual Event, Apr. 25-29*. OpenReview.net, 2022.

[675] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. ERNIE: Enhanced language representation with informative entities. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy, July 2019. Association for Computational Linguistics.

[676] H. Zhong, J. Zhang, Z. Wang, H. Wan, and Z. Chen. Aligning knowledge and text embeddings by entity descriptions. In L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, editors, *Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing (EMNLP'15), Lisbon, Sept. 17-21*, pages 267–272, 2015.

[677] Z. Zhou. Abductive learning: towards bridging machine learning and logical reasoning. *SCIENCE CHINA Information Sciences*, 62(7):76101:1–76101:3, 2019.

[678] H. Zhu, Z. Xiong, S. Magill, and S. Jagannathan. An inductive synthesis framework for verifiable reinforcement learning. In K. S. McKinley and K. Fisher, editors, *Proc. of the 40th ACM SIGPLAN Conf. on Programming Language Design and Implementation (PLDI'19), Phoenix, June 22-26*, pages 686–701. ACM, 2019.

[679] Zhun Yang, A. Ishay, and Joohyung Lee. Neurasp: Embracing neural networks into answer set programming. In C. Bessiere, editor, *Proc.29th Int. Joint Conf. on Artificial Intelligence (IJCAI'20)*, pages 1755–1762, 2020.

[680] L. M. Zouhal and T. Denœux. An evidence-theoretic k-NN rule with parameter optimization. *IEEE Trans. on Systems, Man and Cybernetics C*, 28(2):263–271, 1998.