Abduction and Argumentation for Explainable Machine Learning: A Position Survey*

Antonis Kakas
Department of Computer Science,
University of Cyprus, Cyprus
antonis@ucy.ac.cy

Loizos Michael
Open University of Cyprus &
Research Center on Interactive Media,
Smart Systems, and Emerging Technologies
loizos@ouc.ac.cy

October 27, 2020

Abstract

This paper presents Abduction and Argumentation as two principled forms for reasoning, and fleshes out the fundamental role that they can play within Machine Learning. It reviews the state-of-the-art work over the past few decades on the link of these two reasoning forms with machine learning work, and from this it elaborates on how the explanation-generating role of Abduction and Argumentation makes them naturally-fitting mechanisms for the development of Explainable Machine Learning and AI systems. Abduction contributes towards this goal by facilitating learning through the transformation, preparation, and homogenization of data. Argumentation, as a conservative extension of classical deductive reasoning, offers a flexible prediction and coverage mechanism for learning — an associated target language for learned knowledge — that explicitly acknowledges the need to deal, in the context of learning, with uncertain, incomplete and inconsistent data that are incompatible with any classically-represented logical theory.

1 Introduction

Abduction and Argumentation are two forms of inference where conclusions are drawn according to an underlying theory. Typically, abduction aims to draw an

*A version of this paper will appear in "Dinh Phung, Claude Sammut, and Geoffrey I. Webb (eds). *Encyclopedia of Machine Learning and Data Science, 3rd edition.*, in preparation."

explanation for a set of observations, while argumentation aims to give reasons, or arguments, that support a conclusion against other conflicting conclusions.

Abduction is sometimes described as "deduction in reverse", whereby given a rule "A follows from B" and the observed result "A", we infer that the condition "B" of the rule (may) hold. More generally, in the context of a logic-based setting, given a set of sentences representing a theory T that models a domain of interest, and a sentence representing an observation O, abduction returns a set of sentences representing an abductive explanation H for O, such that:

1. $T \cup H \models O$,
2. $T \cup H$ is consistent,

where $\models$ denotes the deductive (or other) logical entailment relation of the formal logic used in the representation of our theory, and consistency also refers to the corresponding notion in this formal logic. Several abductive explanations can exist for the same observation, and in many cases additional requirements may be imposed, such as for example minimality, for an explanation to be admissible.

Argumentation is concerned with supporting a claim (e.g., a belief or a decision) based on some premises and an argument that links these premises to the claim. An arguments in support of a claim is expected to be acceptable or valid, in the sense of being able to defend itself against all other arguments that are in conflict with it, i.e., counter-arguments that are challenging the supporting argument.

In a formal setting, argumentation takes place within a given argumentation framework[1] $\texttt{AF} = \langle \texttt{Args}, \texttt{Att}, \texttt{Def} \rangle$, where `Args` is a set of arguments, `Att` is a binary attack relation on `Args`, and `Def` is a binary defense (or defeat) relation on `Args`. The attack relation specifies when one argument is a counter-argument to (i.e., opposing or challenging) another argument, while the defense relation captures the notion that an argument is sufficiently strong to defend against another (opposing or challenging) argument. Given an argumentation framework `AF`, a subset of arguments, Δ, is acceptable in `AF`, iff:

(1) Δ is conflict-free; i.e., it does not include arguments that attack each other.

(2) Δ defends against every other subset of arguments A that attacks it[2].

As this definition indicates, the process of argumentation does not in general produce a single supporting argument a_0 for a desired claim, but its aim is to form a coalition, Δ, of a_0 with other arguments so that Δ can defend a_0 against other arguments that would undermine it in some way; e.g., by questioning its premises, or indeed the appropriateness of the argument's link between its premises and the claim (the argument scheme) used to construct a_0.

[1] There are different variations of the definition of an argumentation framework in the literature [14, 44].

[2] More generally, a subset Δ of arguments is acceptable when it can render all its counter-arguments (or counter subsets of arguments) non-acceptable.

Given the above, a claim ϕ is: credulously (or possibly) entailed by AF iff there exists an acceptable coalition Δ under AF that supports ϕ; skeptically (or strongly) entailed by AF iff it is credulously entailed by AF and no other claim ψ that is incompatible (or in conflict) with ϕ (e.g., $\psi = \neg\phi$) is credulously entailed by AF. In effect, deductive entailment is replaced by the more general[3] case of logical reasoning via argumentation. We will see below the significance of this new perspective on reasoning, in relation to the notion of coverage and prediction in logic-based or symbolic learning.

1.1 The Link Between Abduction and Argumentation

Abduction and Argumentation can both be seen as processes for generating explanations either for a given observation as in the case of abduction or for a conclusion (claim or decision) in the case of argumentation. Explanations under abduction are in terms of underlying (theoretical or non-observable) hypotheses, whereas explanations under argumentation are in terms of arguments (among a set of known ones) that provide justified reasons for a conclusion to hold.

Abduction and Argumentation are closely linked. Argumentation can be viewed under the abduction lens as a process of explaining a claim by the set of premises that the supporting argument is based on, linking these to its claim. Furthermore, argumentation can offer a more complete explanation for believing or accepting a claim by providing also reasons against alternative claims, as these form counter-arguments to the chosen claim, which are defended against by the acceptable argument supporting the claim. In other words, these defense arguments form an extension of the explanation for the claim by also giving the reasons why to accept or prefer this claim over other competing alternatives. In the context of abduction, this allows argumentation to be used to provide (and explain) the reasons to prefer one abductive explanation over another abductive explanation for the same observation, as such alternative explanations are typically considered as competing alternatives. Thus, argumentation can help abduction realize its often quoted requirement of "inference to the best explanation".

Another way to understand this link between abduction and argumentation is to notice that abduction can be performed with respect to a theory T that is, in fact, an argumentation framework, i.e., $T =$ AF, and that the entailment relation $\models$ is that of credulous or sceptical entailment under argumentation. With this link abduction can help enhance argumentation in cases where when forming arguments, and defending these against their counter-arguments, it is necessary to make assumptions about missing supporting premises or about other properties that are not known explicitly, but are needed to form a possible defense against counter-arguments. Argumentation then depends on abduction to produce a set of underlying hypotheses that would help the formation of a properly justified argument [32, 61]. These abductive hypotheses become, thus, part of the argument — the reason — for supporting its claim.

[3]Deduction in its classical formulation can be shown to be a limiting case of this form of argumentation-based logic [30].

On the other hand, we can obtain argumentation via abduction by associating to any argument an abducible assumption that carries the burden of the use of the argument such that an argument is attacked by attacking its corresponding assumption, and hence an acceptable set of arguments Δ will correspond to a set of the abducible assumptions linked to the arguments in Δ [15].

2 Motivation & Background: Link to Learning

Reasoning and Learning have a synergistic inter-dependence: learning produces the knowledge that is assumed as given when reasoning; reasoning draws inferences that provide the inductive bias that is assumed as given when learning. How can we exploit this synergy, particularly with the reasoning processes of abduction and argumentation, so that we can enhance the learning process?

The acknowledgement of the existence of this inter-dependence becomes particularly important in the backdrop of a major recent development in Machine Learning and AI: the emergence of the need for explainability in Machine Learning (and Deep Learning, in particular), where decisions taken on the basis of learned models need to be transparent and comprehensible to human users [4, 22], and where neural-symbolic integration can help develop such explainable systems, based both on lower-level sensory data and higher-level cognitive data. Although the detailed operation of a learned hypothesis may be unknown, it should be possible to explain its inferences at a level that is cognitively-compatible with the intended users or consumers of those inferences. Explanations should not require the users to have technological knowledge and should be offered at the high-level conceptual language of the application as used by the application experts and/or users. Abduction and argumentation, as mechanisms that generate explanations, can support directly the realization of this need for Explainable ML, achieving explainability by design of the way they reason over a learned theory.

2.1 Representation and Reasoning for Learned Knowledge

How are we to reason with a learned theory, e.g., to predict the properties of new cases not included in the learning training data? Argumentation (as a conservative extension of deduction in the face of conflicting information) can replace deductive reasoning as a more informative and flexible notion of coverage in learning. The flexibility afforded by argumentation makes it a natural target language for learning, as it acknowledges the default nature of inductively learned knowledge. To demonstrate this natural connection, we consider the celebrated example of Pierce [53] on the color of beans in a bag. Having observed that all beans drawn out of this bag so far are white, and having formed the inductive generalization that "All beans in this bag are white", we face a problem when we draw a non-white bean from the bag. Is the generalization not useful and should be abandoned altogether? If we keep the generalization, how would we reason with it, especially since formal classical deductive logic

would not be appropriate given the observed counter-example of the non-white bean?

As early as in the 18th century, Hume [24] pointed out that inductive generalization that is universal and absolute runs into logical difficulties as we cannot be sure that a future case will not contradict the generalization. Understanding the logic of generalization has come to be known in philosophy as the "the problem of induction" [23]. By the very nature of the task, learning leads to information that in general cannot be absolute, and may contain some element of uncertainty or incompleteness about the underlying knowledge that we are trying to learn. Only in special and limiting cases we arrive at an absolute and complete understanding of what we are trying to learn.

One way to address this foundational question on induction is to consider a target learning language based on argumentative and/or abductive reasoning. Then, inductively-produced rules associating different concepts can be "abductive rules" with a missing element which we need to assume when reasoning with the learned theory. Hence, the learned generalization from the beans in the bag can be represented by "All *normal* beans from this bag are white", where the condition of normality[4] can be abductively assumed for any new bean drawn from the bag. Thus, a new bean will be predicted to be white based on the assumption that it is a normal bean. If indeed it is observed to be white, this would be abductively explained by the assumption that it is a normal bean. If on the other hand, we observe a black bean drawn from the bag, our theory can explain its black color by the assumption that it is not a normal bean.

In time we may be able to learn some properties of these unknown abducible assumptions in our generalizations (learned rules), which would allow us to be more informed about what these missing abductive assumptions are, or deter us from making certain such assumptions; e.g., if the bean feels small then this is not a normal bean in this bag. In particular, it may be possible to start learning other generalizations from the available data that would be usefully integrated together with our earlier generalizations to give an enhanced prediction capability of the integrated learned theory. For example, we may form the inductive generalization that "all small beans from this bag are black", complementing the original generalization that "all beans from this bag are white".

In this way we are naturally led to view the inductive generalizations as providing arguments for the various possible conclusions, rather than strict or absolute rule associations between the concepts involved. For example, the generalization of "all beans from this bag are white" can be interpreted as providing an argument a_1 supporting the claim that a bean is white based on the premise that it was taken from this bag. Similarly, an inductive generalization of "all small beans from this bag are black" provides an argument a_2 supporting the claim that a bean is black based now on the premise that it is a small bean from this bag.

Argumentative reasoning will then juxtapose such arguments in a debate to

[4]Note here assumption of normality refers to a theoretical or non-observable concept, as is the usual case with abductive reasoning [16].

try to reach a conclusion that is supported by an acceptable argument. In this debate, arguments which are (considered) relatively stronger will win, and their claims will be skeptically predicted. In our example above, if a new bean taken from the bag is small then argument a_2 would win the debate provided that we consider a_2 to be stronger than a_1; this relative strength could be naturally justified because the premises of a_2 are more specific than those of a_1.

If no such stronger arguments exist we would typically not be able to reach a clear conclusion, and we would have a dilemma where different and conflicting conclusions could each be supported acceptably by their own arguments. For example, we may have also learned the argument a_3 that wrinkled beans from this bag are green. What would we then predict in a new case where the bean is small and wrinkled? As our learned theory stands, we would have three relevant arguments one claiming white, one black, and the third green as the color of this new bean. The first argument will be defeated by (each of) the other two stronger arguments (due to their specificity over a_1). But between a_2 and a_3 the process of argumentation will not produce a clear winner, and argumentation will offer two credulous predictions, each one accompanied by an explanation.

Although this may not seem to be a completely satisfactory result, the view of learning as a process of producing arguments opens up a new view of reasoning with learned knowledge, where although a final firm decision might not be reached, reasoning is at least able to provide reasons or explanations for concluding one way or another — by presenting (in a suitable way) the arguments that support each of the different possibilities. This could help point towards where we should concentrate any further learning, e.g., the type of further training data that we should collect. Also, potentially, the resolution of such reasoning dilemmas could be offloaded to human experts, embracing a more symbiotic relation between humans and machines working together to reach a decision.

The aforementioned ideas are made precise in our upcoming proposal for a framework on *cognitively-explainable learning* [31], where the use of argumentation as a target learning language is shown to naturally facilitate the formalization of dilemmas in the broader context of building explainable AI systems with learning guarantees. This framework naturally accommodates the consideration of whether the learned argumentation theory is cognitively compatible with the particular human who is consuming the arguments, and who is collaborating with the machine to resolve the dilemmas and draw appropriate inferences.

The feasibility of this symbiosis is supported by decades of work in Cognitive Science and Philosophy, which has confirmed the primacy of argumentation in human reasoning, through a multitude of empirical studies within the area of Cognitive Psychology (e.g., [36]). These empirical studies differ both in the type of experiments carried out, but also in the way that argumentation is formulated for the study. Thus, the underlying position of the primacy of argumentation in human reasoning is solidly confirmed via these different perspectives.

There has also been an extensive study of formulating the many different formal frameworks for non-monotonic — sometimes called commonsense — reasoning in AI, in terms of argumentation, showing the universality of argumentation as a logical framework. With the emphasis (and return) on Human-Centric

and Explainable AI, the use of argumentation as a target language for machine learning could help achieve a seamless integration of AI systems in human lives, by supporting cognitively more comprehensible and understandable learned theories.

2.2 Guiding Learning through Reasons and Explanations

At a technical level, the explanation-generating nature of abduction and argumentation help utilize the current knowledge to facilitate the further and more complete learning. Abduction and argumentation fulfil this function in distinct ways. With abduction explanations provide a form of translation, rationalization, or homogenization of the training data into a common underlying level at which learning is carried out. With argumentation explanations about the reasons why training data are classified in a certain way orients and focuses the learning process.

For abduction, one typically considers that the current knowledge is relatively progressed with a good model, where the incompleteness of the knowledge is isolated to an underlying level that in general is not directly observable or available from the environment. Abduction then generates, in its explanations, new information at this underlying level not hitherto contained in the current theory which can then facilitate the learning of new knowledge in terms of associations between the abducibles. As a case in point, in the context of neural-symbolic integration — an approach that can help develop coherent systems that cope with lower-level sensory data and higher-level cognitive data and enhance, thus, existing deep learning theories with an explainable front-end component — and assuming that the symbolic part of the architecture is mature enough in terms of the knowledge it utilizes, the abductive explanation process of rationalization can help by translating high-level training data into training data at the lower level of a neural module. Therefore, while the neural module might not directly have access to labels that explicitly correspond to its inputs, the process of abduction can translate the high-level label of those inputs into lower-level supervision signals for the neural module.

For argumentation, the current knowledge may not be as developed (it could even be empty), and could still be full of conflicting possibilities for many cases of the training data. Indeed, a problem domain may be inherently one where we cannot (yet) get the same level of isolation of incompleteness of a theory to an underlying level, and we need to tolerate or accommodate the uncertainty of multiple interpretations of cases. By recognizing such dilemma cases (where more than one mutually incompatible prediction would be credulously entailed), argumentation helps to indicate how the current knowledge is to be extended or revised. Knowledge that presents a dilemma on interpretations suggests to the learning process to focus on those training data that would help resolve the dilemma. For example, faced with the dilemma of what color a small wrinkled bean is, learning might seek to identify the conditions under which one argument among those supporting the two choices (of black or green) is stronger.

This means that within an argumentation-based learning framework we want

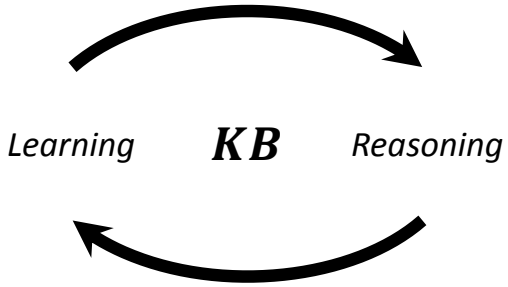

Figure 1: The cycle of reasoning and learning in a process of knowledge development. Learning generates new knowledge KB that feeds into reasoning about new problems. Reasoning about the learning data under the current knowledge KB feeds into the learning process to generate additional knowledge.

to learn not only object-level arguments, but also meta-knowledge in the form of a relative strength or preference between arguments. This meta-knowledge feeds into the defense relation between arguments, which in turn influences the credulous or sceptical predictions that we can draw from the learned theory. Note that such learned meta-level preferences can be conditional on properties of our learning samples, drawing in effect non-linear separation or classification lines in the learning space.

3 Structure of the (Machine) Learning Task

The synergistic inter-dependence between Reasoning and Learning naturally leads to a continuous cycle of interaction between the two processes (see Figure 1), providing for the incremental development of the knowledge KB about the domain of interest. Such an incremental development of the knowledge is necessary when the various knowledge parts (e.g., rules) build on each other, as it ensures that the learning guarantees apply equally for all learned parts [39].

Due to the different nature of abductive and argumentative reasoning, their function in the reasoning-learning cycle is different in its specifics. Both processes, however, have an important role to play in both parts of the cycle: outside learning, they act as explanation-generation mechanisms; and during learning, they utilize their generated explanations to facilitate learning.

Abduction needs prior knowledge KB to operate, and explanations are according to KB. Abductive reasoning influences learning by preparing the learning data and transferring the learning to the underlying level of explanations.

Outside Learning: Abduction uses KB to reason on new cases of interest by providing explanations of the observed status of the new cases.

During Learning: Abductive reasoning *rationalizes*, according to KB, the

training data, thus homogenizing or normalizing the data at an underlying level in KB in a way that can help the inductive learning process. Abductive reasoning also *imputes* missing information in the background attributes/features of the training data, so that under such abductive hypotheses the coverage of a candidate learned theory is improved.

Argumentation reasons with, or analyzes, conflicts in the current knowledge KB. As such, it offers a flexible notion of covering for learning that can help evaluate, understand, or isolate, during the learning process, the inadequacy of the current candidates for the knowledge to be learned, and incrementally revise the learned theory KB through a progressive reduction or resolution of conflicts.

Outside Learning: Argumentation reasons on new cases of interest to predict their status under the learned theory KB, to identify dilemmas among alternative predictions, and to provide reasons supporting each alternative.

During Learning: Argumentative reasoning as a covering notion *interprets* the learning data with reasons for the different possible alternatives for their clarification/interpretation, and makes explicit the dilemmas that exist, according to KB, among those alternatives. These dilemmas help to concentrate on the part of the information that is more relevant and focus the further learning effort on subcases where conflicts continue to exist.

We continue below to expand on the link of the reasoning modes of argumentation and abduction to the learning process, and ground this link on specific approaches as found in the main relevant literature.

3.1 Abduction in Machine Learning

Logic-based abduction generates a set of hypotheses, H, that when added to the given theory T this would logically entail a given set of observations O. This basic formalization as it stands, does not fully capture the explanatory nature of the abductive hypothesis H, in the sense, that it necessarily conveys some reason for why the observations hold. Typically, we want to express the abducive explanations in a way that they convey some "deeper" reason for which the observations must hold according to the theory T, e.g., explanations could be in terms of hypotheses that would causally imply the observations through the given theory. Therefore we would normally specify a "level" in the theory at which the abductive explanations are expressed. One way to do this is simply to specify the vocabulary and language for expressing explanations restricting this to a special preassigned, domain-specific class of sentences called *abducible*.

3.1.1 Abducing Background Knowledge on the Training Data

In the context of Machine Learning the simplest case of this abducible explanation level is the level of the background knowledge associated with the training data. For example, the attribute or feature description of the training examples

may be incomplete for some of these examples. Abductive reasoning can be used to fill in these gaps with informed hypotheses for their value thus helping to prepare the training data for the learning process.

Abductive concept learning (ACL) [33] is a learning framework that allows us to learn from such incomplete information and to later be able to classify new cases that again could be incompletely specified. Under ACL, we learn abductive theories, $\langle T, \mathrm{IC} \rangle$, that contain, together with a set of (logic program) rules, in T, for the concept(s) to be learned, integrity constraints, in IC, in the form of general clauses. These are used to constrain the abductive hypotheses that we can form on the background missing data in classifying new cases according to the learned theory.

The semantics of ACL require that for every positive training example, there must exist in the learned theory an abductive explanation and the collection of all such explanations for all the positive examples must be consistent with each other. For negative training examples, it is required that, within the learned theory, no abductive explanation exists for any of them. We illustrate ACL with a simple example.

Example 1 *Suppose we want to learn the concept father from the following given training examples:*

$$E^+ = \{father(john,\ mary),\ father(david,\ steve)\},$$

$$E^- = \{father(kathy,\ ellen),\ father(john,\ steve)\}.$$

with background knowledge KB given by:

$$KB = \{parent(john,\ mary),\ male(john),\ parent(david,\ steve),\ parent(kathy,\ ellen),\ female(kathy)\}.$$

Notice that the background knowledge that we have on male and female is incomplete and assumptions on these form the abducibles in our problem.
A possible abductive theory, $\langle T,\ IC \rangle$ learned by ACL would consist of

$$T = \{father(X,\ Y) \quad \leftarrow \quad parent(X,\ Y),\ male(X)\},$$

$$IC = \{false \quad \leftarrow \quad male(X),\ female(X)\}.$$

Despite the fact that the background theory is incomplete (in its abducible predicates), ACL can find an appropriate solution to the learning problem by suitably extending the background knowledge and allowing abductive hypotheses, such as that of male(david), in order to cover the positive example of father(david, steve). Note that the learned theory without the integrity constraint that restricts the freedom in drawing abductive hypotheses would not be a solution, because there would exist, in the learned theory, an abductive explanation for the negative example father(kathy, ellen), namely male(kathy), i.e., this negative example would be covered as positive. This explanation is prohibited in the complete theory by the learned constraint together with the fact female(kathy).

The learning algorithm and system for ACL is based on a decomposition of this problem into two subproblems: (1) learning the rules in T together with appropriate explanations in IC for the training examples and (2) learning integrity constraints driven by the explanations generated in the first part. This decomposition allows ACL to be developed by combining the two ILP settings of explanatory (predictive) learning and confirmatory (descriptive) learning. In fact, the first subproblem can be seen as a problem of learning from entailment, while the second subproblem as a problem of learning from interpretations.

An important application of ACL is that of Multiple Predicate Learning (MPL) [55], where each predicate is required to be learned from the incomplete data for the other predicates. Here the abductive reasoning can be used to suitably connect and integrate the learning of the different predicates by generating hypotheses on one predicate on which another predicate depends. This can help to overcome some of the nonlocality difficulties of MPL, such as order-dependence and global consistency of the learned theory.

3.1.2 Abduction as a Case of Learning

In some cases, when our knowledge already contains a detailed description of the problem domain and the incompleteness of its model is isolated in its abducibles, the generation of an abductive explanation/hypothesis from the given set of observations can be considered as a form of learning. This is because the abductive explanation forms a genuine new piece of information that is not already contained in (or derived from) our current knowledge and hence we have learned the underlying reasons or why, things are as observed. In other words, abductive reasoning is equated to a form of learning, sometimes referred to as abductive learning: we have learned the reason(s), according to our current model, for given observational data. In some applications of this form of abductive learning, see, e.g., [57, 62, 63], there can be many such possible abductive explanations specific to the observational data and we then use a probabilistic analysis to extract information out of this multitude of explanations.

In general, though, this form of abductive learning is weak in the sense that its generalization effect, if at all present, is very limited. This is because abduction always needs to refer to some current given theory, from which the explanatory hypotheses are drawn, and thus the generalizing power of abduction is restricted as the basic underlying model of our domain remains unchanged. In other words, the "learned" abductive explanations are already contained in our theory and the possibility to apply these to other cases, other than on the observations that have generated the explanations, is very limited.

3.1.3 Cycle of Abduction and Induction in Learning

To overcome this limitation, we need a synthesis with induction or some other, possibly non-logical, process of learning, where the individual explanations for several different cases of observation are generalized and hence can be applied to genuinely-new cases. Several approaches for synthesizing abduction and induc-

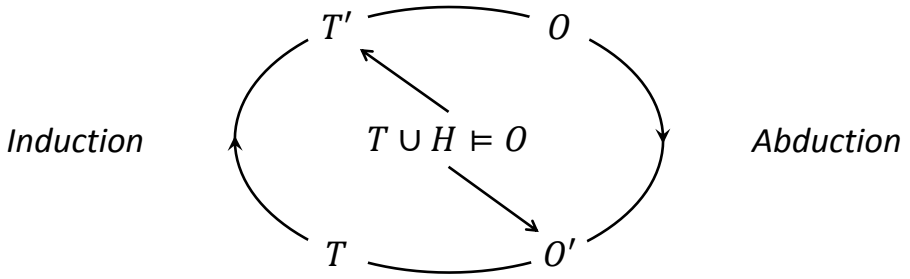

Figure 2: The cycle of abductive and inductive knowledge development [16]. The cycle is governed by the "equation" $T \cup H \models O$, where T is the current knowledge, O the observations triggering theory development, and H the new knowledge generated. On the left-hand side we have induction, its output feeding into the theory T for later use by abduction on the right; the abductive output in turn feeds observational or training data O' for use by induction, and so on.

tion within the process of learning have been developed, e.g., [1, 12, 16, 49, 66]. These approaches aim to develop techniques for knowledge intensive learning with complex background theories. Conceptually, all these proposals fall under the same simple model of actively integrating abduction and induction in a cycle where abduction through its explanations helps to prepare the training data for induction. In general, such a *cycle of integration* of abduction and induction [16] allows the incremental development of the theory T describing our knowledge on a domain (Figure 2).

By treating the given training data for learning as observations, abduction is used to transform (and in some sense normalize) the data to information on the abducible predicates. Then, induction takes this as input and tries to generalize this information to new knowledge on the abducible predicates now treating these as observable predicates for its own purposes. The cycle can then be repeated by adding the learned information on the abducibles back in the model as new partial information on the incomplete abducible predicates. This will affect the abductive explanations of new observations to be used again in a subsequent phase of induction.

A simple example, adapted from [58], that illustrates this cycle of integration of abduction and induction is as follows (see also [12] for further examples):

Example 2 *Suppose that our current model, T, contains the following rule and background facts:*

sad(X) $\leftarrow$ *tired(X), poor(X)*
tired(oli), tired(ale), tired(kr)
academic(oli), academic(ale), academic(kr)
student(oli), lecturer(ale), lecturer(kr)

where the only observable predicate is sad/1.

Given the observations $O = \{sad(ale), sad(kr), not\ sad(oli)\}$ can we improve our model? The incompleteness of our model resides in the predicate poor. This is the only abducible predicate in our model. Using abduction we can explain the observations O via the explanation $E = \{poor(ale), poor(kr), not\ poor(oli)\}$. Subsequently, treating this explanation as training data for inductive generalization we can generalize this to get the rule:

$$poor(X) \quad \leftarrow \quad lecturer(X)$$

thus (partially) defining the abducible predicate poor when we extend our theory with this rule. In other words, we have used abduction to transform the training data from one level, the directly observable level, to training data at another level, the underlying abducible level, where the learning is carried out.

3.1.4 Abduction in ILP Learning

This combination of abduction and induction has been deployed and studied in several ways within the context of ILP. In particular, *inverse entailment* [49] can be seen as a particular case of integration of abductive inference for constructing a "bottom" clause and inductive inference to generalize it. This is realized in Progol 5.0 and applied to several application problems. Similarly, an ILP system called ALECTO [45] integrates a phase of *extraction-case abduction* to transform each case of a training example to an abductive hypothesis with a phase of induction that generalizes these abductive hypotheses. Other ILP frameworks based on the integration of abductive and inductive reasoning can be found in [8, 13, 35].

The development of these frameworks that realize the cycle of integration of abduction and induction prompted the study of the problem of *completeness* for finding any hypotheses H that satisfies the basic task of finding a consistent hypothesis H such that $T \cup H \models O$ for a given theory T, and observations O. Progol was found to be incomplete [66] and several new frameworks of integration of abduction and induction have been proposed such as SOLDR [28], CF-induction [27], and HAIL [58]. In particular, HAIL has demonstrated that one of the main reasons for the incompleteness of Progol is that in its cycle of integration of abduction and induction, it uses a very restricted form of abduction. Lifting some of these restrictions, through the employment of methods from abductive logic programming [29], has allowed HAIL to solve a wider class of problems. HAIL has been extended to a framework, called XHAIL [56], for learning nonmonotonic ILP, allowing it to be applied to learn Event Calculus theories for action description [2] and complex scientific theories for systems biology [59].

3.1.5 Abduction in Neural-Based Learning

In the context of building Explainable AI systems, Abduction can help generate cognitive explanations for Machine Learning predictions. Unlike explanations over logic-based or symbolic learning methods (using, e.g., decision trees), which can be directly obtained from the model [26, 34], explanations over neural or sub-symbolic learning models (using, e.g., deep learning or Bayesian classifiers) are not naturally provided by the model, and need to be obtained by building a parallel symbolic model [25].

In an orthogonal direction, a symbolic module can be built on top of a neural module so that data are fed into the neural module, whose outputs are fed, in turn, into the symbolic module. The latter, then, computes the final outputs, or predictions, of the integrated system, and these predictions are expected to match the high-level labels of the data. In this setting, explanations of the symbolic module take on the role of providing cognitively understandable lower-level labels for the training of the neural module; see, e.g., [10, 64].

It has been shown, in fact, that abduction supports a clean and compositional integration of the neural and symbolic modules, without imposing restrictions on their syntax and semantics [64]. Unlike previous neural-symbolic integration approaches that generally assume that the symbolic module encodes a theory that is effectively differentiable — and, thus, compatible with the typical neural learning process of backpropagation — abduction accommodates any theory for the symbolic module, and utilizes the theory's abductive explanations to compute a loss function that is itself differentiable, even if the theory is not.

It is then easy to see how the particular framework can support an abduction-induction cycle applied on the symbolic module as follows: During the abduction part of the cycle, we assume the symbolic module is fixed, and we abduce labels that we use to train the neural module. During the induction part of the cycle, we assume the neural part is fixed, and we deduce through it data to train the symbolic module. Thus, the cycle iteratively improves how the neural module maps low-level data to high-level data, on which to train the symbolic module.

3.2 Argumentation in Machine Learning

We have argued in Section 2 that argumentation can offer a logical notion of coverage and prediction for learning that would naturally deal with the uncertainty, conflicts and incompleteness that are intrinsic characteristics of learning. Although there are other formulations which share this characteristic, often called non-monotonic logics, the foundational link of argumentation to logic-based learning comes from its advantage that learning can be directly related to building and generalizing arguments from the training data. Indeed, learning can be seen as uncovering the *reasons* for why we observe certain phenomena, i.e., why the given training data is the way it is. We learn therefore arguments that can best support the observed data — that help us classify the data one way instead of another in terms of arguments that support well the given data.

These reasons, or arguments, manifest themselves in various phases of the

learning process, and this manifestation provides a dimension across which relevant work can be categorized. Orthogonally to this categorization, a key aspect that permeates all relevant work is the learning of priorities, or the defense relation, between arguments. Three are the main approaches followed: *(i)* priorities derive from a notion of subsumption on the premises of arguments [9, 21, 51, 52]; *(ii)* priorities are determined by the learning data [11, 38, 39, 40, 50, 54, 60]; or *(iii)* priorities are unavailable or capture domain expert knowledge and are provided externally [3, 5, 6, 17, 18, 19, 41, 42, 48]. Additional considerations regarding a subset of the works reviewed below can be found in [7].

3.2.1 Inputs Extended with Arguments

A first line of work integrates argumentation with learning instances, as *a way to enhance the information that is communicated to the learning process.*

In Argumentation-Based Machine Learning (ABML) [48], a learning instance comprises not only the input and output that specify, respectively, the features and the label of the instance, but also an associated logic-based argument that explains why the particular label is the case as a function of a subset of the features. Thus, upon seeing a small bean with a black color, the learning process also receives the argument "black bean because small bean".

Such side information offers a glimpse of the structure of the target concept that is being learned, which goes beyond what a single labeled learning instance would offer. Unsurprisingly, then, the process of learning benefits greatly in terms of performance, as demonstrated across domains [46, 47, 67].

A similar in flavor approach is also taken in Argumentation-Accelerated Reinforcement Learning (AARL) [17, 18, 19], where the arguments offered provide conditions under which a particular action should be taken in a Reinforcement Learning context. Such arguments can help shape the reward, so that the learning process can then more efficiently learn a good policy.

Extending the basic idea of accompanying learning instances with applicable arguments, Machine Coaching [41, 42] takes the view that these arguments are contextualized on the current state of the learning process, and are not, therefore, available up front. Rather, learning proceeds in an online fashion, where the input of the learning instance is first presented, the learning process makes a prediction on the instance's label, explains how that prediction came about, and only then receives an argument in response to that explanation. The argument itself need not argue with respect to the prediction itself, but could also argue about intermediate concepts that appear in the offered explanation.

When observing a small bean drawn from the bag, for example, the prediction could be that "normal bean because drawn from the bag" and "white bean because normal bean", while the counterargument could be that "not normal bean because small bean". Formal analysis of this protocol shows that efficient learning is indeed facilitated by such counter-arguments, even in cases where efficient learning from input-output instances alone would not be possible.

3.2.2 Inputs Interpreted as Arguments

A second line of work integrates argumentation with the learning process, as *a way to interpret learning instances in case of noisy or partial information.*

Concept Learning as Argumentation (CLA) [3] takes the view that each (noisy) learning instance is an argument that states that for that particular input, the corresponding output is the appropriate instance label. Such arguments are taken to be strong, or preferred, as they derive directly from *observed* information. At the same time, each potential hypothesis among those that could result through learning, offers its own argument on the label of each learning instance, stating that for that particular input, the hypothesis prediction is the appropriate instance label. Such arguments are taken to be weak, as they are premised on the *assumption* that the hypothesis is accurate. Resolving the conflicts between the arguments to compute an acceptable extension effectively amounts to a process of learning, with the extension being the learned outcome.

Restricting attention to only the strong arguments from above, one effectively ends-up with the case-based reasoning (CBR) framework in [9], where each learning instance offers a case in support for its corresponding label. To predict the label of a new learning instance, then, one seeks to find which among the previous cases is closest to the new learning instance in terms of their shared features; such a closest case would then provide an argument for the label of the new learning instance to be the same as the label of that case.

By observing a white bean from the bag and a small black bean from the bag, for example, one ends up with the two arguments "white bean if from the bag" and "black bean if small bean from the bag". Cases (or argument conditions) are not disjoint, and thus multiple cases might be relevant when trying to predict the label of a new learning instance. Argumentation over those cases resolves which cases are to be considered, by offering priority to cases that are more specific: the small bean from the bag, in this example.

Instead of considering individual learning instances as arguments, one can consider arguments derived from subsets of instances. These arguments can, as before, be reasoned with to determine which ones are acceptable, and hence, which ones will be used to support a prediction on a new learning instance. Argumentation for Multi-Agent Inductive Concept Learning (MAICL) [51, 52] approaches this setting by thinking of the subsets of instances as data split among different agents, and by thinking of the corresponding arguments as the hypotheses that were learned by the agents from their given subset of data. In this distributed learning setting, then, argumentation aims to reconcile the different hypotheses that were learned by the agents, so that they can collectively decide the prediction on a new learning instance.

3.2.3 Hypotheses Interpreted as Arguments

Beyond integrating argumentation during the learning process to enhance or interpret learning instances, a third line of work utilizes argumentation after the learning process, as *a way to contrast the learned hypothesis against competing*

learned hypotheses or external alternatives.

Classification enhanced with Argumentation (CleAr) [5, 6] considers a setting where expert opinion or domain knowledge on how a new input should be classified might be available, and these alternatives need to compete against or support the learned hypothesis and each other. Argumentation helps resolve the conflicts that arise, with each argument being assigned a base score, and these scores being used to decide which conclusion prevails.

A similar situation can be observed in the context of unsupervised learning, where one ends up with clusters of training inputs, with each of them effectively providing a competing alternative on how a new input should be classified (i.e., in which cluster it should belong). Argumentation for ART (A-ART) [21] uses argumentation to contrast the competing learned clusters and decide which among them to employ, with priorities given to the various clusters based on a subsumption relation that might exist between them.

MAICL [51, 52], as introduced earlier, can also be seen to fall into this group of approaches, since each agent ends up with a competing learned hypothesis.

3.2.4 Hypotheses Expressed in Argumentation

In the preceding approaches, argumentation was used to provide semantics on how a learned hypothesis interacted with other entities, those being either the learning inputs, or other learned hypotheses or alternative classifiers. The internal structure and semantics of the learned hypothesis was not, however, necessarily specified. In some of the previously described approaches, in fact, hypotheses were, or could be, learned using standard learning algorithms with their associated hypothesis spaces. On the other hand, a body of work has considered argumentation to be the target language for learning, as *a way to specify the representation and reasoning semantics of a learned hypothesis.*

One line of work [50] has considered the problem, and the complexity, of learning an abstract argumentation framework, and hence effectively the attack relation between the arguments, by considering as learning inputs extensions and non-extensions under various argumentation semantics. A number of other works have considered the problem of learning a structured argumentation framework, where the arguments have internal structure, which itself needs to be learned in addition to the attack relation that holds between them.

Work on learning Decision Lists [60] and Exception Lists [38] can be seen to fall in the latter group of works, since such prioritized lists effectively totally order a set of competing conditions as a way to resolve conflicts between their conclusions. In both cases, learning proceeds under Valiant's PAC semantics [65]. Another work [11] approaches the problem under Gold's learning in the limit semantics [20], and shows how learning rules and exceptions can be done iteratively, as a way to systematically enhance the coverage of the learned hypothesis. Priorities between rules effectively correspond to the specificity of rule conditions. Finally, some recent work [54] constructs first an interpretable hypothesis through a standard learning algorithm (random forests, in particular), and subsequently extracts rules from that learned model, which are used as

arguments along with additionally learned conditional priorities between them.

The works above assume that arguments comprise individual rules that directly map their inputs to the supported conclusions. Another body of work has sought to learn argumentation theories with arguments comprising multiple chained rules that map inputs to conclusions through intermediate concepts.

The Never-Ending Rule Discovery (NERD) algorithm [40] offers a heuristic to learn rules in an online fashion where inputs are received in a streaming fashion. In that work, priorities between rules are determined not by the structure of the rules themselves, but are dictated by the learning data, since rules that are found to be sufficiently supported by the data first are given higher priorities. Machine Coaching [41, 42] has taken an alternative approach, where a human coach reacts to the current learned hypothesis and the arguments that are offered in support of a given input, and provides counter-arguments as feedback to enhance the learned hypothesis. Priorities are given based on the recency of the arguments, where more recently-provided counter-arguments defeat earlier ones. Finally, work on Simultaneous Learning and Prediction (SLAP) [37, 39, 43] learns rules in a stratified manner, with priorities between rules, and hence between the resulting arguments, derived from this stratification. Both Machine Coaching [41, 42] and SLAP [37, 39, 43] adopt the PAC semantics [65].

Acknowledgements

This work was supported by funding from the EU's Horizon 2020 Research and Innovation Programme under grant agreements no. 739578 and no. 823783, and from the Government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination, and Development.

References

[1] Hilde Ade and Marc Denecker. AILP Abductive Inductive Logic Programming. In *14th International Joint Conference on Artificial Intelligence (IJCAI)*, IJCAI'95, page 1201–1207. Morgan Kaufmann Publishers Inc., 1995.

[2] Dalal Alrajeh, Oliver Ray, Alessandra Russo, and Sebastián Uchitel. Using Abduction and Induction for Operational Requirements Elaboration. *Journal of Applied Logic*, 7(3):275–288, 2009.

[3] Leila Amgoud and Mathieu Serrurier. Agents that Argue and Explain Classifications. *Autonomous Agents and Multi-Agent Systems*, 16(2):187–209, 2008.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Alejandro Gómez, Daniel Molina, Richard Benjamins, et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Information Fusion*, 58:82–115, 2020.

[5] Lucas Carstens. *Using Argumentation to Improve Classification in Natural Language Problems*. Doctoral Dissertation, Imperial College London, London, U.K., 2016.

[6] Lucas Carstens and Francesca Toni. Improving Out-of-Domain Sentiment Polarity Classification using Argumentation. In *IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1294–1301. IEEE, 2015.

[7] Oana Cocarascu and Francesca Toni. Argumentation for Machine Learning: A Survey. In *6th International Conference on Computational Models of Argument (COMMA)*, pages 219–230, 2016.

[8] Domenico Corapi, Alessandra Russo, and Emil Lupu. Inductive Logic Programming as Abductive Search. In Manuel V. Hermenegildo and Torsten Schaub, editors, *26th International Conference on Logic Programming (ICLP)*, volume 7 of *LIPIcs*, pages 54–63. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2010.

[9] Kristijonas Cyras, Ken Satoh, and Francesca Toni. Abstract Argumentation for Case-Based Reasoning. In *15th International Conference on the Principles of Knowledge Representation and Reasoning (KR)*, 2016.

[10] Wang-Zhou Dai, Qiuling Xu, Yang Yu, and Zhi-Hua Zhou. Bridging Machine Learning and Logical Reasoning by Abductive Learning. In *33rd Conference on Neural Information Processing Systems (NeurIPS)*, pages 2815–2826, 2019.

[11] Yannis Dimopoulos and Antonis Kakas. Learning Non-Monotonic Logic Programs: Learning Exceptions. In *8th European Conference on Machine Learning (ECML)*, pages 122–137. Springer, 1995.

[12] Yannis Dimopoulos and Antonis Kakas. Abduction and Inductive Learning. In Luc De Readt, editor, *Advances in Inductive Logic Programming*, pages 144–171. IOS Press, 1996.

[13] Stanislav Dragiev, Alessandra Russo, Krysia Broda, Mark Law, and Calin-Rares Turliuc. An Abductive-Inductive Algorithm for Probabilistic Inductive Logic Programming. In James Cussens and Alessandra Russo, editors, *26th International Conference on Inductive Logic Programming (ILP)*, volume 1865 of *CEUR Workshop Proceedings*, pages 20–26. CEUR-WS.org, 2016.

[14] Phan Minh Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, 77(2):321 – 357, 1995.

[15] Phan Minh Dung, Robert A. Kowalski, and Francesca Toni. *Argumentation in Artificial Intelligence*, chapter Assumption-Based Argumentation. Springer, Boston, MA, 2009.

[16] Peter Flach and Antonis Kakas. Abductive and Inductive Reasoning: Background and Issues. In P. A. Flach and A. C. Kakas, editors, *Abductive and Inductive Reasoning*, Pure and Applied Logic. Kluwer, 2000.

[17] Yang Gao. *Argumentation Accelerated Reinforcement Learning*. Doctoral Dissertation, Imperial College London, London, U.K., 2015.

[18] Yang Gao and Francesca Toni. Argumentation Accelerated Reinforcement Learning for RoboCup Keepaway-Takeaway. In *2nd International Workshop on Theory and Applications of Formal Argumentation (TAFA)*, pages 79–94. Springer, 2013.

[19] Yang Gao and Francesca Toni. Argumentation Accelerated Reinforcement Learning for Cooperative Multi-Agent Systems. In *21st European Conference on Artificial Intelligence (ECAI)*, pages 333–338, 2014.

[20] E. Mark Gold. Language Identification in the Limit. *Information and Control*, 10(5):447–474, 1967.

[21] Sergio Alejandro Gómez and Carlos Iván Chesnevar. A Hybrid Approach To Pattern Classification Using Neural Networks and Defeasible Argumentation. In *17th International FLAIRS Conference (FLAIRS)*, pages 393–398, 2004.

[22] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A Survey Of Methods For Explaining Black Box Models. *arXiv:1802.01933 [cs.CY]*, 2018.

[23] Leah Henderson. The Problem of Induction. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2020 edition, 2020.

[24] David Hume. *A Treatise of Human Nature*. Oxford University Press, 1739.

[25] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-Based Explanations for Machine Learning Models. In *33rd AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 1511–1519, 2019.

[26] Alexey Ignatiev, Filipe Pereira, Nina Narodytska, and Joao Marques-Silva. A SAT-Based Approach to Learn Explainable Decision Sets. In *9th International Joint Conference on Automated Reasoning (IJCAR)*, pages 627–645. Springer, 2018.

[27] Katsumi Inoue. Induction as Consequence Finding. *Machine Learning*, 55(2):109–135, 2004.

[28] Kimihito Ito and Akihiro Yamamoto. Finding Hypotheses from Examples by Computing the Least Generlisation of Bottom Clauses. In *1st International Conference on Discovery Science (DS)*, pages 303–314. Springer, 1998.

[29] Antonis Kakas, Robert A. Kowalski, and Francesca Toni. Abductive Logic Programming. *Journal of Logic and Computation*, 2(6):719–770, 1992.

[30] Antonis Kakas, Paolo Mancarella, and Francesca Toni. On Argumentation Logic and Propositional Logic. *Studia Logica*, 106(2):237–279, 2018.

[31] Antonis Kakas and Loizos Michael. Cognitively-Explainable Learning. *Manuscript*, in preparation.

[32] Antonis Kakas and Pavlos Moraitis. Argumentation Based Decision Making for Autonomous Agents. In *2nd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, page 883–890. Association for Computing Machinery, 2003.

[33] Antonis Kakas and Fabrizio Riguzzi. Abductive Concept Learning. *New Generation Computing*, 18:243–294, 2000.

[34] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1675–1684, 2016.

[35] Evelina Lamma, Paola Mello, Michela Milano, and Fabrizio Riguzzi. Integrating Induction and Abduction in Logic Programming. *Information Sciences*, 116(1):25–54, 1999.

[36] Hugo Mercier and Dan Sperber. Why Do Humans Reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences*, 34:57–111, 2011.

[37] Loizos Michael. *Autodidactic Learning and Reasoning*. Doctoral Dissertation, Harvard University, Cambridge, Massachusetts, U.S.A., 2008.

[38] Loizos Michael. Causal Learnability. In *22nd International Joint Conference on Artificial Intelligence (IJCAI)*, 2011.

[39] Loizos Michael. Simultaneous Learning and Prediction. In *14th International Conference on the Principles of Knowledge Representation and Reasoning (KR)*, 2014.

[40] Loizos Michael. Cognitive Reasoning and Learning Mechanisms. In *4th International Workshop on Artificial Intelligence and Cognition (AIC)*, pages 2–23, 2016.

[41] Loizos Michael. The Advice Taker 2.0. In *13th International Symposium on Commonsense Reasoning (Commonsense)*, 2017.

[42] Loizos Michael. Machine Coaching. In *IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI)*, pages 80–86, 2019.

[43] Loizos Michael and Leslie G. Valiant. A First Experimental Demonstration of Massive Knowledge Infusion. In *11th International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pages 378–388. AAAI Press, 2008.

[44] Sanjay Modgil. Reasoning about Preferences in Argumentation Frameworks. *Artificial Intelligence*, 173(9):901 – 934, 2009.

[45] Stephen A. Moyle. *An Investigation into Theory Completion Techniques in Inductive Logic Programming*. Doctoral Dissertation, Oxford University Computing Laboratory, University of Oxford, Oxford, U.K., 2000.

[46] Martin Možina, Claudio Giuliano, and Ivan Bratko. Argument Based Machine Learning from Examples and Text. In *1st Asian Conference on Intelligent Information and Database Systems (ACIIDS)*, pages 18–23. IEEE, 2009.

[47] Martin Možina, Jure Žabkar, Trevor Bench-Capon, and Ivan Bratko. Argument Based Machine Learning Applied to Law. *Artificial Intelligence and Law*, 13(1):53–73, 2005.

[48] Martin Možina, Jure Žabkar, and Ivan Bratko. Argument Based Machine Learning. *Artificial Intelligence*, 171(10-15):922–937, 2007.

[49] Stephen H. Muggleton and Christopher H. Bryant. Theory Completion Using Inverse Entailment. In *10th International Workshop on Inductive Logic Programming (ILP)*, pages 130–146. Springer-Verlag, 2000.

[50] Andreas Niskanen, Johannes Wallner, and Matti Järvisalo. Synthesizing Argumentation Frameworks from Examples. *Journal of Artificial Intelligence Research*, 66:503–554, 2019.

[51] Santiago Ontanón, Pilar Dellunde, Lluis Godo, and Enric Plaza. A Defeasible Reasoning Model of Inductive Concept Learning from Examples and Communication. *Artificial Intelligence*, 193:129–148, 2012.

[52] Santiago Ontañón and Enric Plaza. Coordinated Inductive Learning Using Argumentation-Based Communication. *Autonomous Agents and Multi-Agent Systems*, 29(2):266–304, 2015.

[53] Charles S. Peirce. In Charles Hartshorne, Paul Weiss, and Arthur W. Burks, editors, *Collected Papers of Charles Sanders Peirce*. Harvard University Press, 1958.

[54] Nicoletta Prentzas, Andrew Nicolaides, Efthyvoulos Kyriacou, Antonis Kakas, and Constantinos Pattichis. Integrating Machine Learning with Symbolic Reasoning to Build an Explainable AI Model for Stroke Prediction. In *19th IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 817–821. IEEE, 2019.

[55] Luc De Raedt and Nada Lavrač. Multiple Predicate Learning in Two Inductive Logic Programming Settings. *Logic Journal of the IGPL*, 4(2):227–254, 1996.

[56] Oliver Ray. Nonmonotonic Abductive Inductive Learning. *Journal of Applied Logic*, 7(3):329–340, 2009.

[57] Oliver Ray, Athos Antoniades, Antonis C. Kakas, and Ioannis Demetriades. Abductive Logic Programming in the Clinical Management of HIV/AIDS. In Gerhard Brewka, Silvia Coradeschi, Anna Perini, and Paolo Traverso, editors, *17th European Conference on Artificial Intelligence (ECAI)*, volume 141 of *Frontiers in Artificial Intelligence and Applications*, pages 437–441. IOS Press, 2006.

[58] Oliver Ray, Krysia Broda, and Alessandra Russo. Hybrid Abductive Inductive Learning: a Generalisation of Progol. In *13th International Conference on Inductive Logic Programming (ILP)*, volume 2835 of *LNAI*, pages 311–328. Springer Verlag, 2003.

[59] Oliver Ray and Christopher H. Bryant. Inferring the Function of Genes from Synthetic Lethal Mutations. In *2008 International Conference on Complex, Intelligent and Software Intensive Systems (CISIS)*, pages 667–671. IEEE, 2008.

[60] Ronald L. Rivest. Learning Decision Lists. *Machine Learning*, 2(3):229–246, 1987.

[61] Chiaki Sakama. Abduction in Argumentation Frameworks. *Journal of Applied Non-Classical Logics*, 28(2-3):218–239, 2018.

[62] Taisuke Sato and Yoshitaka Kameya. Statistical Abduction with Tabulation. In *Computational Logic: Logic Programming and Beyond*, pages 567–587. Springer, 2002.

[63] Alireza Tamaddoni-Nezhad, Ghazal Afroozi Milani, Alan Raybould, Stephen Muggleton, and David A. Bohan. Construction and Validation of Food Webs Using Logic-Based Machine Learning and Text Mining. In Guy Woodward and David A. Bohan, editors, *Ecological Networks in an Agricultural World*, volume 49 of *Advances in Ecological Research*, pages 225 – 289. Academic Press, 2013.

[64] Efthymia Tsamoura and Loizos Michael. Neural-Symbolic Integration: A Compositional Perspective. *arXiv:2010.11926 [cs.AI]*, 2020.

[65] Leslie G. Valiant. A Theory of the Learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[66] Akihiro Yamamoto. Which Hypotheses Can Be Found with Inverse Entailment? In *7th International Workshop on Inductive Logic Programming (ILP)*, pages 296–308. Springer, 1997. LNAI 1297.

[67] Jure Žabkar, Martin Možina, Jerneja Vidednik, and Ivan Bratko. Argument Based Machine Learning in a Medical Domain. In *1st International Conference on Computational Models of Argument (COMMA)*, volume 144, page 59. IOS Press, 2006.