Argumentation: Reconciling Human and Automated Reasoning

Antonis Kakas[1], Loizos Michael[2], and Francesca Toni[3]

[1] Department of Computer Science, University of Cyprus, Cyprus
antonis@ucy.ac.cy
[2] Open University of Cyprus, Cyprus
loizos@ouc.ac.cy
[3] Department of Computing, Imperial College London, UK
f.toni@imperial.ac.uk

Abstract. We study how using argumentation as an alternative foundation for logic gives a framework in which we can reconcile human and automated reasoning. We analyse this reconciliation between human and automated reasoning at three levels: (1) at the level of classical, strict reasoning on which, till today, automated reasoning and computing are based, (2) at the level of natural or ordinary human level reasoning as studied in cognitive psychology and which artificial intelligence, albeit in its early stages, is endeavouring to automate, and (3) at the level of the recently emerged cognitive computing paradigm where systems are required to be cognitively compatible with human reasoning based on common sense or expert knowledge, machine-learned from unstructured data in corpora over the web or other sources.

1 Introduction

The ever increasing demand for smart machines with ordinary human-level intelligence, attuned to every-day common sense reasoning and simple problem solving, has reignited the debate on the logical nature of human reasoning and how this can be appropriately formalized and automated (e.g. see [14]).

For over half a century now it has been known that the classical view of logic from mathematics is at odds with the nature of human reasoning, as identified in many empirical behaviour studies of cognitive psychology (e.g. see [12] for a relatively recent exposition). Given also that classical logic forms the "calculus of computer science" and the engineering basis for building computing machines, this schism between the two views of logic from mathematics and reasoning from psychology lies at the heart of the problem of automating natural, as opposed to scientific, human reasoning.

In this paper, we first re-consider the foundations of logic from an argumentation perspective, in the spirit of the approach of [28] (see Section 2), and then try to understand how the schism can be reconciled using the argumentation perspective (see Section 3), before finally returning to modern-day cognitive computing (see Section 4) and concluding (see Section 5).

Before analysing how argumentation can be used to reconcile human and automated reasoning, we note that argumentation has been at the heart of logic ever since its inception with Aristotle[4]: logical reasoning according to Aristotle was to follow certain accepted patterns of inference, called *complete deductions*, that were a-priori justified, to find other forms of valid patterns (*syllogisms*) for drawing conclusions. Syllogisms, which in modern logic correspond to derivations in proof theory, are in effect *arguments* for supporting the conclusions that they draw. Moreover, complex arguments can be built from simpler, basic arguments and indeed Aristotle had attempted to show that all *valid arguments* can be reduced to his basic forms of arguments. This reduction task is not easy [2], especially when the given complex argument is obtained through *impossibility*, namely, in modern terms, its conclusion is drawn via a proof by contradiction (or Reductio ad Absurdum (RAA)). This observation is important when logic is formalized through argumentation, as we discuss in Section 2 below.

We also note that *dialectical argumentation* forms a wider context that embraces the conception of logic in Aristotle[5]. A process of dialectic argumentation is based on common beliefs or reputable views from which arguments are built to support conclusions. The important difference with simply demonstrating that a conclusion follows through an argument, is the (extra-logical) additional requirement that the process aims to *convince or persuade*. When logic is formalised as argumentation, both views of logic (as validity and as dialectics aiming at persuading) play an important role, as we discuss in Section 2 below.

2 Argumentation and Logical Reasoning

From the point of view of argumentation, reasoning can be seen as a double process: constructing *arguments* linking premises and conclusions that we want to draw as well as *defend* these arguments against (possible or proposed) counter-arguments generated from an *attack* relation between arguments.

Arguments are built from constituent parts (e.g. private or common beliefs or reputable views) that are combined in some way to form a structure supporting desired positions. In a logical setting the process of constructing an argument for a desired position can be associated to a logical proof for the position. Hence, given some language, $\mathcal{L}$, for logical sentences together with a notion of deduction, $\vdash$, drawn from some set of inference rules, an argument can be constructed by choosing some set, Δ, of sentences in $\mathcal{L}$ — the argument's *premises* — that under the application of (some of the) inference rules deduce, via $\vdash$, a *conclusion* that is identified with the desired position. Arguments may also draw intermediate conclusions (combined, as dictated by the inference rules of $\vdash$, to give the main conclusion) and thus combine sub-arguments within a single argument structure.

A logical language would typically also contain some notion of contradiction or inconsistency (denoted by $\bot$), which can be used, in conjunction with the inference rules, to provide the notion of counter-argument/the attack relation

[4] A concise introduction to Aristotle and logic can be found in [2].

[5] A wider exposition of logic and dialectical argumentation can be found, e.g., in [46].

between arguments. In the simplest case, two arguments would attack each other when they, or some of their sub-arguments, support contradictory positions [4].

Several methods have been proposed in the literature on argumentation in artificial intelligence (e.g. see [3, 44] for overviews) to determine when arguments may be deemed to defend against counter-arguments generated from attacks. We will informally illustrate below how this notion of defence against counter-arguments can be formalized using some standard semantical notions given for abstract argumentation [10, 24]. Within the resulting formalization, logical entailment of a sentence ϕ from a theory T, traditionally defined in Propositional Logic (PL) in terms of truth of ϕ in all models of T, can be given, informally put, by the statement **"there is an acceptable argument (from T) that supports ϕ, and any argument that supports $\neg\phi$ is not acceptable"**. Furthermore, this argumentation-based entailment is also defined when the given theory T may be classically inconsistent, i.e. without classical models. In the case of classically inconsistent theories, PL trivializes, whereas the argumentation semantics continues to differentiate between sentences that are entailed and sentences that are not entailed, as we will see below.

For the illustration, let us consider a simple example, where we take as the underlying language $\mathcal{L}$ a standard language of classical PL allowing to represent the following rules (and all other knowledge/beliefs presented later in this section):

- "normally, a seller who delivers on time is trustworthy";
- "normally, a seller who delivers the wrong merchandize is not trustworthy".

These rules can be represented[6] in PL by the following sentences (for some seller in the (finite) domain of discourse):

$$timely\_delivery \rightarrow trusted \tag{1}$$
$$wrong\_delivery \rightarrow \neg trusted \tag{2}$$

Given additional information about a delivery by the seller we can build arguments for and against trusting the seller. For example if we observe that the seller delivers on time

$$timely\_delivery \tag{3}$$

then we can build argument $\mathbf{A}_1$ with premises the sentences (1) and (3) and conclusion that the seller should be trusted. Moreover, if the seller has made a wrong delivery

$$wrong\_delivery \tag{4}$$

[6] As we will discuss later, under the argumentation-based reformulation of PL the implication connective used here does not need to be interpreted as classical material implication. Rather, an implication $A \rightarrow B$ may be interpreted, informally, as "given A we have an argument for B".

then we can build argument $\mathbf{A}_2$ with premises the sentences (2) and (4) and conclusion that the seller should not be trusted. These two arguments are counter-arguments against (or attack) each other when indeed we have both pieces of information (3) and (4). Furthermore, if we also have the additional rule

– "if the seller is trusted then one can place large orders",

represented as

$$trusted \rightarrow large\_orders \tag{5}$$

then we can build argument $\mathbf{A}_3$ with premises (3), (1) and (5) and conclusion that large orders can be placed with the seller in question. Note that argument $\mathbf{A}_2$ is still a counter-argument against $\mathbf{A}_3$, despite the fact that the (final) conclusions that they support are not contradictory, as $\mathbf{A}_2$ *undercuts* $\mathbf{A}_3$ on $trusted$ on which $\mathbf{A}_3$ depends. In other words, $\mathbf{A}_2$ attacks $\mathbf{A}_3$ because it attacks a sub-argument of $\mathbf{A}_3$, namely the sub-argument $\mathbf{A}_1$.

Given arguments and attacks between them, we can determine which arguments are (dialectically) *acceptable*, e.g. in the spirit of *admissibility* in abstract argumentation [10], and define notions of *credulous entailment* and *sceptical entailment* from the theory from which arguments and attacks are built in terms of this acceptability. In the earlier example, given a theory T amounting to sentences (1)–(5) above:

– arguments $\mathbf{A}_1$ and $\mathbf{A}_2$, as defined above, are both (individually) acceptable; for example, $\mathbf{A}_1$ is acceptable as it does not attack itself (as sentences (1) and (3) are consistent) and $\mathbf{A}_1$ *defends* itself against the attack by $\mathbf{A}_2$;
– $trusted$ and $\neg trusted$ are both credulously entailed by T, since the arguments $\mathbf{A}_1$ and $\mathbf{A}_2$ are both acceptable;
– neither $trusted$ nor $\neg trusted$ are sceptically entailed by T, since their (respective) negation is credulously entailed by T.

Note that it may be useful in some cases to use *hypotheses* as premises of arguments whose conclusions are credulously entailed if the hypotheses are dialectically legitimate and the arguments using them as premises are acceptable. For example, it may be desirable to hypothesize

$$\neg large\_orders \tag{6}$$

to form an argument $\mathbf{A}_4$ with premises the sentences (2) and (4) from the given theory and hypothesis (6) to support the conclusion $\neg large\_orders$ while also including, within the premises of the argument, its defences against attacks. The argument $\mathbf{A}_4$ is acceptable because it can defend against all attacks. For example, the attacks against $\mathbf{A}_4$ by argument $\mathbf{A}_1$ and by argument $\mathbf{A}_3$ are both defended against by $\mathbf{A}_4$ since they are defended against by the sub-argument $\mathbf{A}_2$ of $\mathbf{A}_4$. This is depicted in figure 1 below, showing the dialectic nature of the argumentation semantics of acceptability. Here, $\mathbf{A}_3$ disputes the hypothesis $\neg large\_orders$ in $\mathbf{A}_4$, and the sub-argument $\mathbf{A}_1$ of $\mathbf{A}_3$ disputes the conclusion

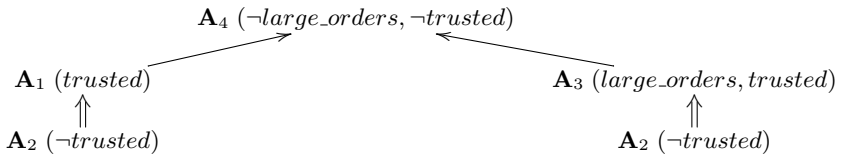

Fig. 1. Dialectic process of argumentation: the sub-argument $\mathbf{A}_2$ of $\mathbf{A}_4$ defends (represented by $\Uparrow$) against the attacks (represented by $\uparrow$) by arguments $\mathbf{A}_1$ and $\mathbf{A}_3$ against argument $\mathbf{A}_4$. For all arguments, we indicate in brackets the sentences of contention.

$\neg trusted$ of sub-argument $\mathbf{A}_2$ of $\mathbf{A}_4$. Notice that the defence by $\mathbf{A}_2$ against $\mathbf{A}_3$ does not come simply from the symmetric incompatibility of $\neg large\_orders$ on $\mathbf{A}_3$. This can be justified via an implicit priority of premises from the given theory over hypotheses: arguments solely supported by the given theory are stronger than those supported by hypotheses. Notice that priorities on premises can also be given explicitly [27], e.g. we may consider the premise $wrong\_delivery \rightarrow \neg trusted$ to be stronger than $late\_delivery \rightarrow trusted$.[7] This would have the effect that $\mathbf{A}_2$ would attack $\mathbf{A}_1$ but not vice versa and hence only $\neg trusted$ would be credulously entailed by the above theory.

It is easy to see that credulous entailment of a sentence corresponds to a form of satisfiability, i.e. if a sentence is credulously entailed then there exists a maximal consistent sub-theory of T with which the sentence is also consistent. More significantly, for appropriate definitions of arguments, attack and defence [26, 28], sceptical entailment corresponds to classical logical entailment PL, if the theory is consistent.

When working with classically inconsistent theories, PL can use the RAA inference rule (or proof by contradiction) to derive, by means of an *indirect proof*, any sentence based on a subset of contradictory premises in the theory. For this reason, most works in logic-based argumentation, e.g. [4], impose the restriction that the premises of arguments be classically consistent. Instead, the approach that we advocate [28] redefines logic in terms of argumentation (rather than retaining classical logic and building arguments based on this) and ascribes a different nature to the RAA rule, of an argumentative nature rather than as a building block in the construction of arguments, restricted instead to be *direct proofs*, e.g. using a Natural Deduction system [15] but without any use of the RAA rule.

Separating out the RAA rule and excluding this from being one of the primary means to construct arguments gives rise to (a form of) **Argumentation Logic** (AL) [27, 28] and allows us to overcome the technical difficulties of working with inconsistent premises, that Aristotle had to face too. AL offers a semantically equivalent reformulation of classical PL in the realm of classically consistent

[7] This is analogous to assigning higher strength to associations that refer to a more specific subclass in inheritance reasoning such as the famous AI example of "penguins not fly" considered as a stronger property association than "birds fly".

theories of premises that smoothly extends into the inconsistent realm without trivializing as PL does. We will informally describe AL here and illustrate it by means of examples (the technical details are not essential here and can be found in [27, 28]). AL defines a form of sceptical entailment, as indicated earlier, but in terms of recursive notions of acceptability and non-acceptability of arguments [24] rather than admissibility as in [10]. Arguments are identified with their premises, which may be drawn from a given theory as well as hypothesized, as discussed earlier, and constructed with a notion of *direct derivation* based on a subset, $\vdash_{DD}$, of standard inference rules in Natural Deduction that does not contain the RAA inference rule. The important technical aspect of AL is that the inferences from the RAA rule are recovered semantically through the notion of non-acceptability of arguments. Informally, showing that a hypothesis, ϕ, is inconsistent is replaced by showing that arguments which contain ϕ are non-acceptable and hence such arguments cannot lead to the entailment of any sentences. This is different from the classical interpretation of the RAA rule which in addition leads to the (sceptical) entailment of $\neg\phi$. This additional step is not present in AL and this absence gives AL flexibility to reason with contradictory information.

In AL, the notions of acceptability and non-acceptability of arguments are defined dialectically in terms of notions of *attack* and *defence* amongst arguments, where defence is defined as a restricted type of attack, ensuring asymmetry (as discussed earlier).

Acceptability is defined as a relative notion between arguments, i.e. that an argument a is acceptable with respect to another argument a_0, informally meaning that if we a-priori accept a_0 then a is an acceptable argument. The argument a_0 helps to render a acceptable. For example, when we want to adopt an argument this could be used in its defence against counter-arguments and analogously an argument could render itself non-acceptable by rendering one of its counter-arguments acceptable. The informal definition of this acceptability notion of "*a is acceptable with respect to a_0*" is that for every argument, b, attacking a there is an argument, d, that defends against b and this defending argument"*d is acceptable with respect to $a \cup a_0$*". The defending argument d in effect renders the attacking argument non-acceptable. Then, non-acceptability is defined as the classical negative dual of acceptability. Finally, a sentence (seen as an argument) is (sceptically) entailed from a theory if it is acceptable with respect to the empty argument and its negation (seen again as an argument) is not acceptable with respect to the empty argument. To illustrate this let us consider the additional statement:

– "normally, a seller that delivers late is not trustworthy",

represented as

$$\neg timely\_delivery \rightarrow \neg trusted \tag{7}$$

and, further, assume that we have also learned (by analyzing our sale records) that, for our particular seller,

– "normally, when the delivery is on time the wrong item is delivered",

represented as

$$timely\_delivery \rightarrow wrong\_delivery \tag{8}$$

Consider now the theory T consisting of sentences (1), (2), (7) and (8). This is classically consistent and it classically entails $\neg trusted$. Let us see how in AL we would derive that *trusted* is non-acceptable, that is needed for AL to (sceptically) entail $\neg trusted$. The non-acceptability of *trusted* can be determined by considering the argument $\mathbf{B}_1$ with premises T and hypotheses $\{trusted\}$, and the dialectical process of argumentation depicted in figure 2.[8]

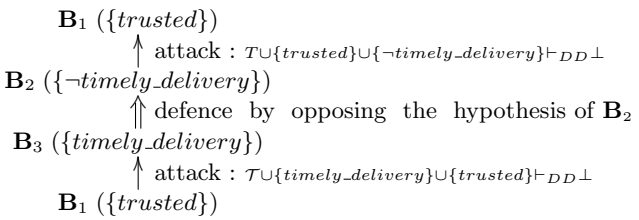

Fig. 2. Determining the non-acceptability of *trusted*, with respect to the empty argument, given $T = \{(1), (2), (7), (8)\}$, in order to determine classical entailment of *trusted* from T. All arguments have T as their premises and hypotheses as indicated in brackets.

The figure shows that there is an attacking argument ($\mathbf{B}_2$) for which there is no acceptable defence as the only possible defence ($\mathbf{B}_3$) against this attack is attacked by the proposed argument ($\mathbf{B}_1$): the defence is rendered non-acceptable by the proposed argument and hence the proposed argument is not acceptable.

Note that the example theory T is classically consistent and the above dialectic proof of the non-acceptability of *trusted* can be shown [28] to be connected to the proof in Natural Deduction that this hypothesis is inconsistent with the theory via a nested use of the RAA inference rule. In general, AL captures classical propositional reasoning in dialectic argumentation terms when the premises are classically consistent, but also that same dialectical process extends into the case of classically inconsistent theories as a form of para-consistent reasoning. This is achieved by linking RAA in PL with the recognition of a particular class of non-acceptable arguments, namely arguments that are self-defeating.

The following results of AL [28] are important for our discussion in this paper:

– AL can be chosen to be exactly equivalent to classical PL when the theory is classically consistent

[8] For simplicity we assume here that $\vdash_{DD}$ contains only the Modus Ponens inference rule.

– AL is a weakening of classical PL that does not trivialize when the theory is classically inconsistent
– AL avoids logical paradoxes such as the Barber of Seville paradox
– The classical interpretation of the implication connective is not forced onto AL; when the antecedent is false both the implication and its negation are acceptable
– AL corresponds to a special form of Natural Deduction where the application of the RAA rule is restricted

But perhaps the most important property of AL, shared by most argumentation frameworks proposed in AI over the last 25 years, is its dialectic nature of reasoning that gives a natural form of *bounded reasoning* by considering incrementally different attacking arguments as they arise in the case of reasoning at hand through new (evidential) information that is brought in the theory or to the attention of the reasoner: a form of *on-demand reasoning* which, as we will see in the next section, is supported by psychological evidence [35].

Reconciliation - level 1: Given the above results on AL we have reached our **first level of reconciliation of human and automated reasoning**. Assuming that argumentation and argumentative dialectic reasoning is closer to human reasoning than strict classical logic and thus human reasoning can be automated through argumentation, it is important to know that an argumentation perspective is not a radical deviation of the status quo in automated reasoning. Indeed, the whole notion of computation and its formalization and automation rests on the foundation of PL, Boolean algebra and von Neumann computer architectures. Adopting the new foundation of argumentation for logic and reasoning does not abandon the existing frameworks of automated reasoning but co-exists with the foundation of computation through classical logic as the "calculus of computer science".

From now on we will use the term AL to refer, generically, to the re-formulation of logical reasoning via argumentation.

3 Argumentation and Human Reasoning

The aim of this section is to present some of the main features of human reasoning coming out from empirical studies in the Psychology of Logic (Reasoning) and to examine how formal argumentation in AI, and in particular AL, conforms with these features. We will also overview some recent work from Psychology that gives direct evidence for argumentation in human reasoning. Putting these results together with the previous section we will argue that human reasoning can be formalized well through argumentation in a way that facilitates its automation for building artificial intelligent systems.

Over the last century, a large amount of research has been carried out in the area of Psychology of Reasoning with the results suggesting that human reasoning is failing, in comparison with strict mathematical or classical logic, at simple logical tasks, committing mistakes in probabilistic reasoning and being subject to irrational biases in decision making [12, 22]. Earlier on, Stoerring [49]

showed empirically that humans perform with significant variations in successfully drawing conclusions under different classical logic syllogisms. One way to interpret this difference, as discussed in [48], is to recognize that humans do not reason according to the classical entailment of truth in all possible models of the premises but rather they reason in an intended interpretation of the premises. The authors of [48] go further to propose that syllogistic classical reasoning is addressed by human subjects as constructing a suitable situation model much like when humans are processing or comprehending a narrative.

Humans do well when using Modus Ponens with implications. But they do not fair well in using Modus Tollens, which also indicates that they have difficulty in reasoning by contradiction (i.e. in applying the classical RAA rule). On the other hand, humans recognize that the falsification of an implication comes through a case where the condition (antecedent) of an implication holds yet the conclusion of the implication does not hold [5]. This indicates that, although humans also recognize when the condition (antecedent) of an implication is false the status of the conclusion is irrelevant, they do not do this by recognizing that the implication is trivially satisfied, as is the standard view in classical logic, but rather by recognizing that the implication cannot be argued against, i.e. it is not possible to falsify the implication in a situation where the antecedent does not hold.

Most work in the Psychology of Reasoning then points to the observation that human reasoning differs from reasoning in classical logic and different interpretations and theories on the nature of human reasoning have been proposed. On the one hand, there are proposals that stay very close to the mathematical and strict form of logical reasoning, such as the proposal of "The Psychology of Proof" theory [45], which proposes a psychological version of a proof system for human reasoning in the style of Natural Deduction. Despite many criticisms, see e.g. [19] for a thorough and critical review of this theory, it shows a necessary departure from the proof systems of classical logic but more importantly it implicitly indicates that human reasoning is linked to argumentation since proof systems as that of Natural Deduction have a natural argumentative interpretation, as we have seen in Section 2. Other proposals, e.g. [7], abandon completely any logical form for human reasoning treating it as the application of specialized procedures, invoked naturally depending on the situation that people find themselves.

Importantly, the study of the Psychology of Syllogisms [18, 21, 20] proposes that humans use mental models to guide them into drawing inferences. Humans, in general, construct an intended mental model which captures the logic of the situation at hand and do not consider alternative models, as in classical reasoning, so as to ensure the "absolute and universal" validity of the inference. As mentioned above this has also being pointed out in [47, 48]. In a modern manifestation of this position, based on the nature of Computational Logic in Artificial Intelligence, such as how Logic Programming can be applied to problems of human intelligence, the recent book [29] argues that building structures like mental

models is a useful way to capture various features of human reasoning, not least its defeasible nature, which as we will see below is central in our discussion.

But building mental models can be seen as building arguments from the available evidence currently at hand and general premises of common sense knowledge that people have acquired. Then, as argued also in [29], the mental model approach to deduction can be reconciled with the view of reasoning through inference rules. Here we go one step further and argue that the process of building the mental models is the dialectic process of argumentation, based on acceptable arguments: we will discuss an example below in Section 3.2.

Summarizing, the work in the Psychology of Reasoning has exposed the following *salient features* of human reasoning given here from a modern view of computational logic in Artificial Intelligence:

- Human reasoning is able to handle contradictory information without trivializing. There is no inconsistent state of the human mind that would make it incapable of drawing any inference.
- Human reasoning is defeasible. Conclusions drawn can be retracted in the face of new information. Knowledge on which human reasoning is based is not absolute.
- An implication (or a rule) expresses only an association between its conditions and conclusion, not a necessity. Other properties of an implication that are implicit in the classical logical interpretation are not prominent in human reasoning.

Finally, we separate out some recent work from the Psychology of Reasoning which provides explicit evidence for the argumentative nature of human reasoning. In [35] the authors have proposed, based on a variety of empirical psychological experiments, that human reasoning is a process where humans provide reasons to accept (or decline) a conclusion that was "raised" by some incoming inference of the human brain. The main function of reasoning is to lay out these inferences in detail and form possible arguments that will produce the final conclusion and therefore, through the process of reasoning, people will be able to exchange arguments for assessing new claims: the process of reasoning is a process of argumentation.

What characterizes the process of reasoning proper is the awareness, not just of a conclusion, but of an argument that justifies accepting that conclusion. To validate their argumentation theory for reasoning, the psychologists have carried out experiments to test how humans form, evaluate and use arguments. One important conclusion of their study is the fact that humans will come up with "solid" arguments when they are in an environment where they are motivated to do so, i.e. in an environment where their position is challenged. Otherwise, if not challenged the arguments produced could be rather naive. But once counter-arguments or opposing positions are put forward people will produce better and well justified arguments for their position by finding counter-arguments (i.e. defences) to the challengers. For example, in experiments where mock jurors were asked to reach a verdict and then were presented with an alternative one

it was observed that almost all of them were able to find counter arguments against the alternative verdict (very quickly) strengthening the arguments for their original verdict.

This indicates that automating human reasoning through argumentation can follow a model of computation that has an *"on-demand"* incremental nature. This will be well suited in a resource bounded problem environment and more generally for cognitive systems that we will consider in the next section.

3.1 Human Reasoning and Argumentation in AI

The Psychology of Reasoning had influenced AI in its effort to automate human reasoning. From the early stages of AI an approach based on *production rules* was developed, influenced by the psychological findings on the nature of implications in human reasoning. Cognitive architectures [1, 30] for systems were proposed whose baseline computation is given by the application of production rules and following onto their conclusions when these were drawn. Despite their relative success (and their re-emergence today in the new era of Cognitive Computing that we will examine in Section 4) these systems were considered to be lacking a proper formal foundation. For example, how was the firing of a production rule to be interpreted? On the one hand when its conditions hold it must fire to give its conclusion and yet the conclusion could be at odds with conclusions of other production rules that have also fired. Attempts to provide formal foundations for these rules have been made [11].

To address these shortcomings but also independently motivated by the desire to have a clear formal semantics of intelligent computation in AI, new formal logical frameworks, beyond classical logic, were proposed (starting with the seminal works [33, 34], with several other later approaches) and two areas of AI were established for this: (1) Non-monotonic Reasoning and (2) Belief Revision. The emphasis was on formal logics within which conclusions could be withdrawn or revised when additional information rendered the theory (classically) inconsistent. Furthermore, the need for non-monotonic logics was also advocated directly by psychologists from an early stage (see e.g. [47]) by recognizing these logics as reasoning in intended models. But despite the wealth of theoretical results the variety of and differences amongst the various approaches has prevented a clear consensus on the question of what is the formal nature (if any) of human reasoning.

Nevertheless, with the introduction of formal argumentation in the early 1990s, to capture in particular the non-monotonic reasoning of negation as failure in Logic Programming [23], it was shown (e.g. in [6, 10]) that argumentation can capture different approaches of non-monotonic reasoning and thus provide a uniform framework for it. Argumentation is now used as the underlying framework to study and develop solutions for different types of problems in AI [3, 44]. In particular, it forms the foundation for a variety of problems in multi-agent systems (see the workshop series at `http://www.mit.edu/~irahwan/argmas`) where agents need to exhibit human-like intelligence, but with emphasis on autonomy and adaptability rather than human-like reasoning and problem solv-

ing. Recently, *argument mining* (see [32] for an overview), aims to provide an automatic way of analysing, as argumentation frameworks of some sort, human debates in social media, even by identifying relations not explicit in text [50].

All these studies of argumentation in AI show that argumentation has the capacity to address the salient features of human reasoning that the empirical studies of psychology have pointed out. Argumentative reasoning is a natural form of reasoning with contradictory information by supporting arguments for conflicting conclusions, such that conclusions are withdrawn when new stronger arguments come into play. Argumentation gives a form of reasoning in an intended mental model, the model corresponding to the conclusions that are supported by the stronger arguments available, which is naturally defeasible as this model can change in the face of new information.

3.2 Argumentation for Story Comprehension

The argumentative nature of human reasoning is quite pronounced in the central task of comprehension of a given situation or narrative. Humans try to comprehend situations they are faced with through a mental model constructed by inferring information that is not explicitly present in the situation, yet it follows from it, explaining why things happened as they did, linking seemingly unconnected events, and predicting how things will further evolve, by developing arguments in each case. To construct such an intended comprehension model it is necessary to be able to draw inferences relating to how information changes over time (e.g. the story line) whether these are about the state of the physical or mental (e.g. intentions of protagonists in a story) world in the narrative.

Several works [13, 16, 25, 39] have shown that argumentation can be used to capture such reasoning about actions and change, introduced in AI by the seminal works of the Situation and Event Calculi. Argumentation can address the three central problems associated with this, namely the frame, ramification and qualification problems, in a natural way by capturing this aspect of human reasoning in terms of persistence, causal and default world property arguments and a natural attacking relation between these different types of arguments. Grounding these types of arguments on information that is explicitly given in the narrative, we can build arguments for and against drawing a certain conclusion at a certain time point or situation in the world.

Recently, combining this argumentation approach to reasoning about actions and change with empirical knowhow and theoretical models from Cognitive Psychology, it was possible to show that argumentation can successfully capture many of the salient features of narrative text comprehension by human readers, as exposed through many years of empirical and subsequent theoretical studies from Cognitive Psychology. In particular, the mental comprehension model that is built by human readers of stories corresponds to the grounded extension of a corresponding argumentation framework whose arguments are grounded on the explicit information in the story [8]. An associated automated comprehension system [9], shows how these different basic forms of reasoning involved, such

as the persistence of information across time, the change brought about by actions, and the blocking of change when this violates default properties, can be automated through argumentation.

Consider, for example, a scenario where we hear Alice saying to her colleague: "Bob delivered the first car we ordered on time, but he brought us the wrong model. This is a big problem. We should cancel the other order with Bob.". Upon hearing Alice, we seek to bridge her utterances. We may reason that Alice might be invoking the following premises $timely\_delivery \rightarrow trusted$ and $wrong\_delivery \rightarrow \neg trusted$ to build arguments for and against trusting Bob. From Alice's second sentence, that there is a problem, we might infer that the second premise and argument built from it is stronger for Alice, and that despite the conflicting inferences of the two arguments, Alice is led to infer that Bob is not trusted. We may further consider that Bob's trustworthiness related to the cancelling of the second order through the premise $\neg trusted \rightarrow \neg large\_order$, and thus we make sense of the situation. We would as easily make sense of the situation had Alice said "This is a minor problem. We should keep the other order with Bob.", realizing that the preference on Alice's premises and arguments built from them are reversed. For more information the reader is referred to the web site of the STAR system for Story Comprehension through Argumentation at `http://cognition.ouc.ac.cy/star`.

Despite the fact that further work is needed to address fully the important problems of coherence and abstraction in narrative text comprehension the fact remains that argumentation can provide a basis on which to build further these important features of the comprehension model.

Reconciliation - level 2: Argumentation logic and argumentative dialectic reasoning is closer to human reasoning than strict classical logic capturing well the features that crucially characterize and distinguish human level natural intelligence from mathematical and scientific reasoning. Argumentation treats all human knowledge as inherently defeasible building from these arguments to support conclusions. Argumentation is able to reconcile all approaches to date in AI for automating human reasoning under one umbrella that conforms to the empirical observations on the nature of human reasoning in Psychology. As such Argumentation Logic and argumentation in general provides a solid foundation for automating human reasoning in AI able to accommodate reasoning with inconsistent premises and revision of conclusions.

4 Argumentation for Cognitive Computing

We now present how argumentation can form a foundational basis for the new paradigm of Cognitive Computing and for the development of cognitive systems, which, as we have argued in the Introduction, forms the renewed impetus for the need to automate human reasoning.

Cognitive systems, as studied for example in [31, 42] and the IBM Watson machine, however they are defined, have two important characteristics. First, they are *cognitively compatible* to humans in the sense that there exists a level of

communication between a cognitive system and a human user analogous to the communication between humans: cognitive systems need to be able to articulate their results or decisions to human users while at the same time they need to comprehend instructions, e.g. personal preferences of human users, not as programming commands but as overall requirements on the desired solutions of a problem. At the end the aim of computation in these systems is to *persuade* a human user that they have a *good or acceptable solution* to a problem rather than to compute an objectively absolute correct solution to the problem.

In effect, what we are asking is for these systems to be able to reason like humans and to interact with humans in the form of a dialectic argumentative dialogue. Given the discussion of the previous sections where we have argued that human reasoning can be formalized in terms of argumentation and that formal argumentation from AI captures well the human dialectic process of argumentation, we believe that argumentation can form the basis for the *Reasoning API* of cognitive systems.

But argumentation can have a much deeper role in the development of the field of Cognitive Computing. This new paradigm needs a new foundational notion of computation that classical logic cannot provide but which formal argumentation is well suited for. This is linked to the second main characteristic of cognitive systems, the fact that these systems depend and operate on *knowledge acquired from unstructured data* normally through some form of machine learning, again analogously to the situation that we have in humans who learn over time the knowledge on which they base their reasoning. For example, cognitive home or work assistants, analogous to personal assistants, need to have common sense knowledge so that they can humanly comprehend their environment, which would normally also include human users, from the explicit (but sparse) information that this gives to them, which is sufficient for a human to comprehend the environment. This is much like story comprehension where this needs the integration of the explicit information in the narrative with common sense background world knowledge.

How is then common sense knowledge to be acquired? The field of cognitive computing implicitly assumes that this, or any other form of expert knowledge whether this is "data science" knowledge, or knowledge from "data analytics", or even "mined arguments", on which a cognitive system will be built, would be obtained incrementally through a process of (largely) autonomous learning from unstructured data. Learning frameworks that integrate symbolic learning and reasoning have shown that this can be done in a manner that the learned knowledge is guaranteed to draw inferences that are probably approximately correct [37], while exploiting raw text (e.g. from the Web) as their input [36, 38], and accommodating the interaction of different pieces of learned knowledge in supporting a drawn inference [40, 43, 51].

Such learned knowledge cannot form strict absolute knowledge as it could in a didactic supervised form of learning. It would be defeasible knowledge that holds for the most part. It expresses typical not absolute relationships between concepts, where these links are dependent on the various contexts and sub-contexts

of the problem domain. Hence the form of such knowledge can be naturally associated to argumentation: learned knowledge from unstructured data forms the basis, the premises, for building arguments. In philosophical terms, the inductive syllogism, as Aristotle calls this process of acquiring first principles from empirical experience, cannot produce absolute knowledge: an inductively produced implication $A \rightarrow B$ does not formally express the "necessity" of B when A is known to hold but an argument for B thus making B "probable" in the current case, as the philosopher David Hume [17] suggests. Recent work seeks to extend the aforementioned learning results to the case of learning such implications [41].

We note that this formalization of such learned knowledge in terms of argumentation does not apply only to common sense knowledge but also to expert knowledge acquired from unstructured data. Hence when IBM Watson is applied to oncology in health applications the scientific knowledge learned provides the basis for acceptable arguments that the machine can build in its task to give medical recommendations. These arguments provide a way for the Watson machine to explain the structure and quality of its recommendation to the human doctor as one human doctor would do to another doctor.

Reconciliation - level 3: The new foundation for logic and reasoning of argumentation can provide the basis for a more flexible paradigm of computing on which to build systems that are cognitively compatible to humans. Such cognitive systems are built by exploiting common sense knowledge or expert knowledge learned from unstructured data whose formal understanding can naturally be captured by argumentation logic.

5 Conclusions

Drawing from past and recent work in Psychology and AI, we have argued that an alternative view of logic as a framework for argumentation is closer to natural human reasoning than classical logic. As such an argumentation-based formulation of logical reasoning offers a way to reconcile and bridge human and automated reasoning.

Within argumentation the notion of classical logical entailment can be formulated in a way that conforms with an open ended dialectic process of argumentation offering new possibilities and features that are attuned to human reasoning. Important such features included handling conflicting information and a form of bounded rationality of "on-demand" reasoning where arguments are defended against counter-arguments that are grounded on the evidential information of the case at hand rather than on hypothetical situations that might possibly arise.

Argumentation Logic allows us to give a different interpretation to symbolic knowledge that on the one hand has the flexibility of human reasoning while at the same time offers a suitable target language for learning from unstructured data. This makes argumentation a suitable logical foundation for the emerging paradigm of Cognitive Computing. In this new paradigm the abstract notion of computation can be formalized as the construction of acceptable arguments supporting solutions of the problems. These solutions are not necessarily globally

and objectively optimal but rather solutions that are locally convincing and persuasive, according to the expectation of the human user, both for the case where the problem is a common sense task or a task in an expert domain.

To some extent our proposal constitutes a return back to early AI where cognitive psychology had a strong influence on the field. The crucial difference is that now we are advocating argumentation as a new logical foundation for this human-level AI, away from the, albeit many times implicit, assumption that classical logic can form the basis for automating human reasoning, as it does for all other conventional computing problems within the realm of scientific and engineering problems. In fact, AI, by insisting on a classical logic foundation for intelligence, took a turn towards problems that fell in this engineering realm with an emphasis on "super intelligence" beyond the level of common sense intelligence one would find ordinarily in humans. But for cognitive systems where knowledge is typically incomplete and inconsistent from a classical logic perspective, a radical change in the formal foundations of intelligent computation is needed. We have argued that this can be given through a reformulation of logic through argumentation as the primary notion of logical reasoning.

Given the need for a new logical foundation for cognitive computation one might ask whether this would also need its own architecture of computers on which to be realized, in the same way that the Von Neumann architecture is linked to classical (Boolean) logic. What would such an architecture be? Could it be a connectionist architecture where the threshold activation of signal propagation is linked to argument construction under inputs for and against the argument, thus giving argumentation a final reconciliation role between symbolic and connectionist approaches to automating human reasoning?

References

1. J. R. Anderson and C. Lebiere. *The Atomic Components of Thought.* Lawrence Erlbaum Associates, 1998.
2. J. Barnes. *The Cambridge Companion to Aristotle.* Cambridge University Press, 1995.
3. T. J. M. Bench-Capon and P. E. Dunne. Argumentation in Artificial Intelligence. *Artificial Intelligence*, 171(10-15):619–641, 2007.
4. P. Besnard and A. Hunter. *Elements of Argumentation.* The MIT Press, 2008.
5. E. W. Beth and J. Piaget. *Mathematical Epistemology and Psychology.* Dordrecht: Reidel, 1966.
6. A. Bondarenko, F. Toni, and R. Kowalski. An Assumption-based Framework for Non-monotonic Reasoning. In *Proceedings of the 2nd International Workshop on Logic Programming and Non-monotonic Reasoning (LPNMR)*, pages 171–189, 1993.
7. P. Cheng and K. Holyoak. Pragmatic Reasoning Schemas. *Cognitive Psychology*, 17:391–416, 1985.
8. I.-A. Diakidoy, A., L. Michael, and R. Miller. Story Comprehension through Argumentation. In *Proceedings of the 5th International Conference on Computational Models of Argument (COMMA)*, pages 31–42, 2014.

9. I.-A. Diakidoy, A. Kakas, L. Michael, and R. Miller. STAR: A System of Argumentation for Story Comprehension and Beyond. In *Proceedings of the 12th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense)*, 2015.
10. P. M. Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-person Games. *Artificial Intelligence*, 77:321–357, 1995.
11. P. M. Dung and P. Mancarella. Production Systems with Negation as Failure. *IEEE Transactions on Knowledge and Data Engineering*, 14(2):336–352, 2002.
12. J. S. Evans. Logic and Human Reasoning: An Assessment of the Deduction Paradigm. *Psychological Bulletin*, 128(6):978–96, 2002.
13. N. Y. Foo and Q. B. Vo. Reasoning about Action: An Argumentation-theoretic Approach. *Artificial Intelligence Research*, 24:465–518, 2005.
14. U. Furbach and C. Schon, editors. *Proceedings of the Workshop on Bridging the Gap between Human and Automated Reasoning — A workshop of the 25th International Conference on Automated Deduction (CADE-25), Berlin, Germany, August 1, 2015*, volume 1412 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.
15. G. Gentzen. Unstersuchungenuber das Logische Schliessen. *Mathematische Zeitschrift*, 39:176–210, 1935. English translation in M. Szabo (ed.), The Collected Papers of Gehard Gentzen, North Holland, Amsterdam, 1969.
16. E. Hadjisoteriou and A. Kakas. Reasoning about Actions and Change in Argumentation. *Argument and Computation*, 2016. doi:10.1080/19462166.2015.1123774.
17. D. Hume. *A Treatise of Human Nature*. Oxford: Clarendon Press, 1888. Originally published 1739-40. Edited by L. A. Selby Bigge.
18. P. Johnson-Laird. *Mental Models*. Cambridge University Press, 1983.
19. P. Johnson-Laird. Rules and Illusions: A Critical Study of Rips's The Psychology of Proof. *Minds and Machines*, 7(3):387–407, 1997.
20. P. Johnson-Laird and R. M. J. Byrne. *Deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991.
21. P. Johnson-Laird and M. Steedman. The Psychology of Syllogisms. *Cognitive Psychology*, 10:64–99, 1978.
22. D. Kahneman and A. Tversky. Subjective Probability: A Judgment of Representativeness. *Cognitive Psychology*, 3(3):430 – 454, 1972.
23. A. Kakas, R. Kowalski, and F. Toni. Abductive Logic Programming. *Journal of Logic and Computation*, 2(6):719–770, 1992.
24. A. Kakas and P. Mancarella. On the Semantics of Abstract Argumentation. *Journal of Logic and Computation*, 23(5):991–1015, 2013.
25. A. Kakas, R. Miller, and F. Toni. An Argumentation Framework of Reasoning about Actions and Change. In *Proceedings of the 5th International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR)*, pages 78–91, 1999.
26. A. Kakas, F. Toni, and P. Mancarella. Argumentation and Propositional Logic. In *Proceedings of the 9th Panhellenic Logic Symposium (PLS)*, 2013.
27. A. Kakas, F. Toni, and P. Mancarella. Argumentation for Propositional Logic and Nonmonotonic Reasoning. In *Proceedings of the 11th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense)*, 2013.
28. A. Kakas, F. Toni, and P. Mancarella. Argumentation Logic. In *Proceedings of the 5th International Conference on Computational Models of Argument (COMMA)*, pages 12–27, 2014.
29. R. Kowalski. *Computational Logic and Human Thinking: How to Be Artificially Intelligent*. Cambridge University Press, New York, NY, USA, 2011.

30. J. E. Laird. *The Soar Cognitive Architecture*. MIT Press, 2012.
31. P. Langley. The Cognitive Systems Paradigm. *Advances in Cognitive Systems*, 1:3–13, 2012.
32. M. Lippi and P. Torroni. Argumentation Mining: State of the Art and Emerging Trends. *ACM Transactions on Internet Technology*, 16(2):10, 2016.
33. J. McCarthy. Programs with Common Sense. In *Semantic Information Processing*, pages 403–418. MIT Press, 1968.
34. J. McCarthy. Circumscription — A Form of Non-monotonic Reasoning. *Artificial Intelligence*, 13(1):27 – 39, 1980.
35. H. Mercier and D. Sperber. Why do Humans Reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences*, 34:57–74, 4 2011.
36. L. Michael. Reading Between the Lines. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1525–1530, 2009.
37. L. Michael. Partial Observability and Learnability. *Artificial Intelligence*, 174(11):639–669, 2010.
38. L. Michael. Machines with WebSense. In *Proceedings of the 11th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense)*, 2013.
39. L. Michael. Story Understanding... Calculemus! In *Proceedings of the 11th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense)*, 2013.
40. L. Michael. Simultaneous Learning and Prediction. In *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pages 348–357, 2014.
41. L. Michael. Cognitive Reasoning and Learning Mechanisms. In *Proceedings of the 4th International Workshop on Artificial Intelligence and Cognition (AIC)*, 2016.
42. L. Michael, A. Kakas, R. Miller, and G. Turán. Cognitive Programming. In *Proceedings of the 3rd International Workshop on Artificial Intelligence and Cognition (AIC)*, pages 3–18, 2015.
43. L. Michael and L. G. Valiant. A First Experimental Demonstration of Massive Knowledge Infusion. In *Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pages 378–389, 2008.
44. I. Rahwan and G. R. Simari. *Argumentation in Artificial Intelligence*. Springer Publishing Company, 1st edition, 2009.
45. L. J. Rip. *The Psychology of Proof: Deductive Reasoning in Human Thinking*. Cambridge MA: MIT Press, 1994.
46. M. Shenefelt and H. White. *If A, Then B: How Logic Shaped the World*. Columbia University Press, 2013.
47. K. Stenning and M. van Lambalgen. *Human Reasoning and Cognitive Science*. MIT Press, 2008.
48. K. Stenning and M. van Lambalgen. Reasoning, Logic, and Psychology. *WIREs Cognitive Science*, 2:5:555–567, 2010.
49. G. Stoerring. Experimentelle Untersuchungen ueber einfache Schlussprozesse. *Archiv fuer die gesammte Psychologie*, 11:1–127, 1908.
50. F. Toni and P. Torroni. Bottom-Up Argumentation. In *Proceedings of the 1st International Workshop on Theories and Applications of Formal Argumentation (TAFA)*, pages 249–262, 2012.
51. L. G. Valiant. Robust Logics. *Artificial Intelligence*, 117(2):231–253, 2000.