On the Meta-Logic of Arguments

Michael Wooldridge* Peter McBurney* Simon Parsons†

*Dept of Computer Science
University of Liverpool
Liverpool L69 7ZF, U.K.
mjw,peter@csc.liv.ac.uk

†Dept of Computer and Information Science
Brooklyn College, CUNY
Brooklyn, 11210 NY, USA
parsons@sci.brooklyn.cuny.edu

ABSTRACT

Argumentation has received steadily increasing attention in the multi-agent systems community over the past decade, with particular interest in the use of argument models from the informal logic community. The *formalisation* of such argument systems is a necessary step if they are to be successfully deployed, and their properties rigorously understood. However, there is as yet no widely accepted approach to the formalisation of argument systems. In this paper, we take as our starting point the view that arguments and dialogues are inherently *meta-logical*, and that any proper formalisation of argument must embrace this aspect of their nature. For example, a statement that serves as a justification of an argument is is statement *about* an argument: the argument for which the justification serves must itself be referred to in the justification. From this starting position, we develop a formalisation of arguments using a hierarchical first-order meta-logic, in which statements in successively higher tiers of the argumentation hierarchy refer to statements further down the hierarchy. This enables us to give a clean formal separation between object-level statements, arguments made about these object level statements, and statements about arguments.

Categories and Subject Descriptors

I.2.11 [**Distributed artificial intelligence**]: multiagent systems; I.2.4 [**Knowledge Representation Formalisms and Methods**]: predicate logic

General Terms

theory, languages

Keywords

multi-agent systems, argumentation, knowledge representation, meta-level reasoning

AAMAS'05, July 25-29, 2005, Utrecht, Netherlands.

1. INTRODUCTION

Argumentation has received steadily increasing attention in the multi-agent systems community over the past decade, with particular interest in the use of argument models from the informal logic community such as that of Walton and Krabbe [19, 24]. The *formalisation* of such argument systems is a necessary step if they are to be successfully deployed, and their properties rigorously understood. Most argument systems can be classified according to whether the arguments they consider are structured, typically logical entities (e.g., [2, 11, 12, 13]), or atomic, abstract entities (in the sense of Dung's abstract argument model [7, 1]). However, although some research has considered the links between these different types of systems [3], no one model is universally accepted, and both the abstract and logical argumentation paradigms have well-known problems as a model of rational argument [18].

In this paper, we focus on a logic-based view of arguments [13]. We take as our starting point the view that arguments and dialogues are inherently *meta-logical* processes. By this, we mean that the arguments made by protagonists in a debate must *refer* to each other. This is because arguments are not just about which states of affairs exist in the world, or how objects in the world stand in relation to one-another. If this were the case, then dialogues would be impoverished indeed, essentially restricted to asserting the truth or falsity of statements. We believe that rational argumentation also involves putting forward *arguments about arguments*, and it is in this sense that they are *meta-logical.* For example, a statement that serves as a justification of an argument is a statement *about* an argument: the argument for which the justification serves must itself be *referred to* in the justification.

One of our main aims in this paper is to put this idea of meta-argument on the map of argumentation research. But we also hope to show how a meta-logical treatment of argument can clarify some apparently difficult issues in the formalisation of argument. Our basic approach involves developing a *hierarchical* formalisation of logic-based arguments. That is, we construct a (well-founded) tower Δ_0, Δ_1, $\ldots$ of arguments, where arguments, statements, and positions at a level n in the hierarchy may refer to arguments and statements at levels m, for $0 \leq m < n$. In the bottom tier Δ_0 of the hierarchy are *object level* statements about the domain of discourse. The apparatus we use for formalising such an argument system is a *hierarchical first-order meta-logic*, a type of first-order logic in which individual terms in the logic can refer to terms in another language (cf. Konolige's first-order formalisation of knowledge and action [10]). This formalisation enables us to give a clean formal separation between object-level statements, arguments made about these object level statements, and statements about arguments.

The remainder of the paper is structured as follows. First, in the following section, we give a motivation and informal introduction to the framework. In section 3, we present a proof-of-concept formalisation of our approach using hierarchical meta-logic, and in section 4, we present some conclusions. Our work makes two key contributions to the theory of argumentation. First, and perhaps most importantly, we motivate and establish the notion of meta-argumentation as an issue in its own right, and present a first formalisation of this process. Although meta-*languages* have been used in the formalisation of dialectical systems [20], to the best of our knowledge we are the first to use a *meta-logic* in this way. Our second contribution is to show how a number of different approaches to argumentation may be uniformly combined within the meta-logic framework: in particular, the logic-based approaches of [2, 13], the abstract argumentation framework of Dung [7], and Bench-Capon's value-based argumentation framework [1]. Note that the integration of abstract argument frameworks and logic-based frameworks is possible only *because* we adopt a meta-logical perspective: the integration involves stating and reasoning about relations over logical formulae, which cannot be achieved without some meta-logical apparatus.

2. A HIERARCHICAL SYSTEM OF ARGUMENTS

Before proceeding to the formal details of our approach, we present some more detailed motivation for it. As noted in the introduction, our key motivation is the following observation:

$$\text{Argumentation and formal dialogue is necessarily a meta-logical process.} \qquad (*)$$

This seems incontrovertible: even the most superficial study of argumentation and formal dialogue indicates that, not only are arguments made about object-level statements, they are also made *about* arguments. In such cases, an argument is made which *refers* to another argument. Moreover, there are clearly also cases where the level of referral goes even deeper: where arguments refer to arguments that refer to arguments. Perhaps the paradigm examples of such meta-argumentation would be in a courtroom setting, where an advocate objects to an argument of the opposing advocate, or where a judge rules an argument inadmissible. Here, the arguments being put forward refer to arguments made about the domain of discourse, but are clearly not actually about the domain of discourse itself.

If one accepts the validity of (*), then it is natural to view argument as taking place at a number of levels. At the lowest level, we do not really have arguments at all – we have statements about the domain of discourse. At the next level in the argumentation hierarchy, we have arguments themselves: these are statements about the object-level statements, and so on. Of course, in any attempt to formalise such a model of arguments, we must define the composition of each level of the hierarchy. There are many choices to be made here – particularly at higher levels of the hierarchy – and we are in no position to give a canonical view. In this paper, we set out and work with a 3-tier hierarchy, as illustrated in Figure 1. Throughout the remainder of the paper, we will denote these levels of the hierarchy by Δ_0, Δ_1, etc., with Δ_0 always being the *lowest* level of the hierarchy. The tiers of the hierarchy are as follows:

Δ_0 *The Object Level:* This tier of the hierarchy does not actually contain arguments at all. It consists of statements about the domain of discourse, and in particular defines the interrelationships between the entities in the domain of discourse. In a legal setting (which is perhaps the paradigm example of a domain for formal argument and discourse), we can think of Δ_0 as consisting of the established facts of the case, (such as evidence that may be introduced), as well as non-logical axioms about the domain.

Δ_1 *Ground Arguments:* Arguments exist for the first time as first class entities in this tier of the hierarchy. Δ_1 defines what constitutes an argument: in the model of argument that we use, an argument consists of a conclusion and some supporting statements, with a notion of logical consequence between them [2, 13]. By contrast, in Toulmin's scheme an argument is more complex, consisting of a claim (e.g., "John is old") , a warrant (e.g., "over 70 is old") with associated backing (e.g., some demographic data), and some data (e.g., "John is 78") [22]. Note that the hierarchical meta-logic approach itself is consistent with both such models of argument, and indeed many others; but we find it convenient to work with the logical model. Since we can refer to arguments in this tier in the hierarchy, we can also capture relationships *between* arguments here. For example, the canonical notion of one argument *attacking* another is a relation between arguments [7], and cannot therefore be present at any lower tier of the hierarchy. Although "attack" is one relation that may exist between arguments, it is of course not the only one: since the object level Δ_0 will often contain inconsistencies, the notion of attack will often not be enough to obtain a useful coherent view. We therefore use Bench-Capon's notion of *value*-based argument, which overlays attack with *values* that the argument appeals to, and hence makes it possible to choose between arguments on the basis of the values they represent [1].

Δ_2 *Meta-Arguments:* Notice that at the Δ_1 tier of the hierarchy, we can make statements that are about object-level statements, (e.g., we can assert that a particular structure represents an acceptable argument) but we cannot directly refer to the *process* by which an argument is established. That is, in Δ_1 we cannot say that "we can establish that a is an argument using axioms T". Hence properties of arguments that involve referring to the axioms or procedures via which we in fact establish that they are arguments cannot be captured in Δ_1. However, such properties *can* be captured in Δ_2. In particular, the main construction used in Δ_2 is that of an argument referring to an argument. To illustrate the value of this, we will show how we can distinguish in Δ_2 between "classical" $\mathcal{L}_1$ arguments (in which the full technical apparatus of classical logic proof can be used to establish a conclusion), and *intuitionistic* $\mathcal{L}_1$ arguments, where a more restrained (and

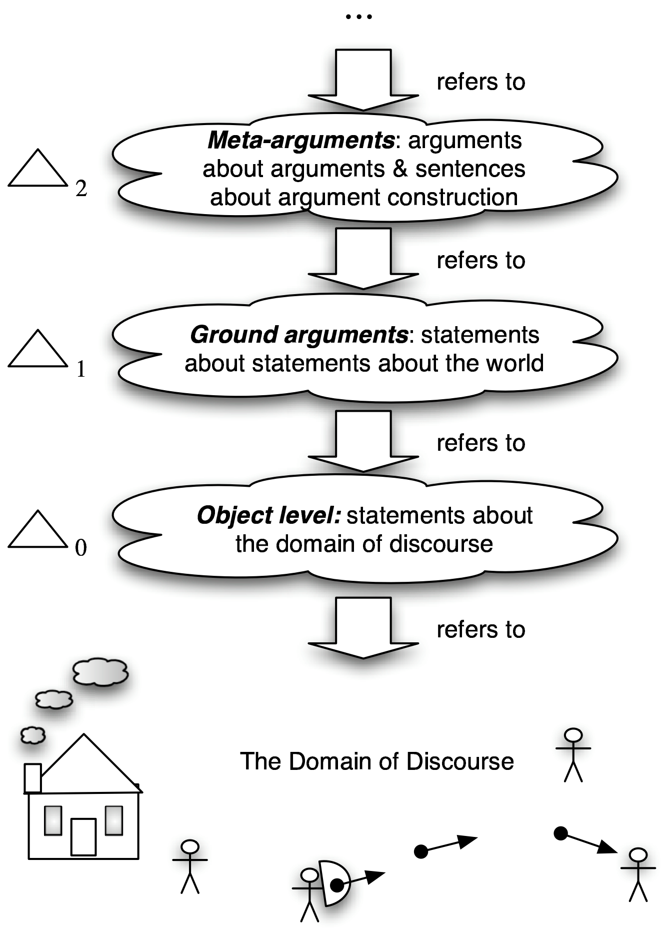
...
refers to
Δ_2
Meta-arguments: arguments about arguments & sentences about argument construction
refers to
Δ_1
Ground arguments: statements about statements about the world
refers to
Δ_0
Object level: statements about the domain of discourse
refers to
The Domain of Discourse

Figure 1: A hierarchy of arguments.

some would argue more realistic) notion of proof is used [6].

Of course, there is no reason why this hierarchy should not be continued: the same logical apparatus we use can essentially be copied into layers further up the hierarchy, permitting arguments about arguments about arguments ... as desired. Where argumentation is used in human settings, this is exactly what seems to happen: consider an argument that takes place between advocates in a court of law, and then (further up the hierarchy), arguments made by the judge about these arguments, and then potentially arguments made in a supreme court about the arguments made by the judge in the lower court. To cleanly (and properly) capture this kind of setting, it seems to us that our hierarchical approach is not only appropriate, but perhaps essential. However, for the purposes of this paper, we will restrict our attention to the three layers indicated here.

3. THE FORMAL FRAMEWORK

Meta-level reasoning (reasoning about reasoning) has a venerable history in artificial intelligence, and *logical* approaches to meta-level reasoning have been widely studied, with a range of approaches developed and evaluated (see [8, pp.239–262] and [16, 5] for reviews). For our purposes, the most suitable formalism to adopt is a *first-order meta-logic* [8, Chapter 10]. Viewed in the most abstract way, a first-order meta-logic is simply a first-order logic whose domain (the set of entities that may be referred to in the language) includes sentences of another language (the *object* language). An important distinction is made between meta-languages that can refer to themselves (i.e., languages whose domain contains the set of sentences of the language itself), which are usually called *self-referential*, and those where this is not possible. Self-referential languages tend to be rather complex and intricate systems to deal with: first because when one assumes even seemingly innocent and innocuous axioms they tend to become inconsistent, and second because they allow one to express paradoxical statements such as the "liar" paradox [15, 23]. *First-order hierarchical meta-languages* provide a somewhat more stable logical foundation [23]. The basic idea of such languages is that we define a (well-founded) tower of languages $\mathcal{L}_0 - \mathcal{L}_1 - \cdots$, such that the domain of $\mathcal{L}_0$ is the set of entities in the domain of discourse, and the domain of each language $\mathcal{L}_u$ for $u > 0$ contains the set of formulae of language $\mathcal{L}_v$ for $0 \leq v < u$, but contains no sentences from languages $\mathcal{L}_w$ for $w \geq u$. In this way, we have the ability to make statements about statements about statements ... to some arbitrary level of depth, but because our languages are strictly hierarchical (can only refer to sentences of languages further down the hierarchy), self-reference (and all the logical problems it entails) is not possible. Hierarchical meta-languages have been used as the basis of several formalisms for reasoning about action (see, e.g., [10]) and recently the approach of using meta-language predicates in place of modal operators for referring to (for example) what an agent knows or believes has undergone something of a revival (see, e.g., [9]). We will present our formalisation of each tier in the hierarchy in turn, starting with Δ_0.

Note that we do not give a syntax and semantics for each language $\mathcal{L}_i$, as these are available elsewhere in the literature (e.g., [10, 23]). We will assume that the languages contain the conventional logical connectives of negation ("$\neg$"), disjunction ("$\vee$"), conjunction ("$\wedge$"), implication ("$\rightarrow$"), and bi-conditional ("$\leftrightarrow$"), the usual apparatus of first-order quantification ("$\forall$" "$\exists$"), functional terms, equality, and logical constants for truth ("**true**") and falsity ("**false**"). Moreover, for each language $\mathcal{L}_i$ we assume a logical consequence relation $\models_{\mathcal{L}_i}$. Technically, each level Δ_i in our hierarchy will constitute a *theory* in the language $\mathcal{L}_i$.

3.1 The Object/Domain Level: Δ_0

We can understand Δ_0 as stating the basic "facts" of the argumentation domain, and the non-logical axioms associated with it[1]. We often refer to Δ_0 as the object-level, or domain theory. Thus, in the domain theory Δ_0, we define all the properties about the argumentation domain that may be admitted into the discourse. For simplicity of exposition here, we will assume that these are expressed using propositional logic, although of course there is no reason in principle why one should not use a richer language. Formally, Δ_0 will be a set of formula expressed in propositional logic.

[1] By "non-logical" axioms, we mean axioms or rules which refer specifically to the domain at hand, and which are not valid according to the semantics of the logic.

Example 1. *Here is an example domain theory:*

$$\Delta_0 = \{p, t, p \to q, (q \vee r) \to \neg s, t \to \neg p\}.$$

3.2 Arguments About the Domain: Δ_1

Let us now move one step up the hierarchy. At level 1 in the hierarchy, we define our basic model of arguments: what constitutes an acceptable argument according to the underlying system of argument that we are interested in. In line with [2, 13], we consider an argument with respect to a domain theory Δ_0 as a pair $\langle \varphi, \Gamma \rangle$, such that:

1. $\varphi \in \Delta_0$ is an $\mathcal{L}_0$-formula known as the *conclusion* of the argument and $\Gamma \subseteq \Delta_0$ is a set of $\mathcal{L}_0$-formulae known as the *support*;
2. Γ is consistent (i.e., not $\Gamma \models_{\mathcal{L}_0}$ **false**);
3. φ logically follows from Γ (i.e., $\Gamma \models_{\mathcal{L}_0} \varphi$); and
4. there is no subset Γ' of Γ satisfying (2) and (3).

We now formalise this in our hierarchical logic framework. We must first put in place some conventions. First, recall that the domain of language $\mathcal{L}_1$ contains the expressions of $\mathcal{L}_0$. We assume that, for each primitive $\mathcal{L}_0$ expression e, there is a corresponding $\mathcal{L}_0$ term e'. $\mathcal{L}_0$ terms denoting compound object-language formulae are constructed using the meta-language functions *and*, *or*, *not*, and so on. Thus *or* is an $\mathcal{L}_1$ functional term which takes two arguments, each of which is an $\mathcal{L}_1$ term denoting an $\mathcal{L}_0$ formula: the function returns the $\mathcal{L}_0$ sentence corresponding to the disjunction of its arguments. For example, the $\mathcal{L}_0$ formula

$$p \to (q \vee r)$$

is denoted by the $\mathcal{L}_0$ term

$$\mathit{imp}(p', \mathit{or}(q', r')).$$

Since this construction is somewhat cumbersome, we follow standard practice and use *sense quotes* (sometimes called Frege quotes or Gödel quotes) as abbreviations:

$$\begin{array}{lcl} \ulcorner \neg p \urcorner & \hat{=} & \mathit{not}(p') \\ \ulcorner p \vee q \urcorner & \hat{=} & \mathit{or}(p', q') \\ \text{etc.} & & \end{array}$$

We will also assume that we have terms in $\mathcal{L}_1$ that stand for *sets* of $\mathcal{L}_0$ formulae. To build sets formally, we use an $\mathcal{L}_1$ constant $\emptyset$, which denotes the empty set of $\mathcal{L}_0$ formulae, and unary function $\mathit{set}(f)$, which takes an $\mathcal{L}_1$ term denoting an $\mathcal{L}_0$ formula, and returns the singleton set of $\mathcal{L}_0$ formulae containing the formula denoted by f. Finally, we use a binary function $\mathit{union}(T_1, T_2)$, which takes as arguments two $\mathcal{L}_1$ terms, each of which denotes a set of $\mathcal{L}_0$ formulae, and returns the set of $\mathcal{L}_0$ formulae corresponding to the union of these two sets. To make this somewhat more readable, we will write

$$\{\ulcorner \varphi_1 \urcorner, \ulcorner \varphi_2 \urcorner, \ldots, \ulcorner \varphi_k \urcorner\}$$

as an abbreviation for the following, somewhat more cumbersome $\mathcal{L}_0$ term:

$$\mathit{union}(\mathit{set}(\ulcorner \varphi_1 \urcorner), \mathit{union}(\mathit{set}(\ulcorner \varphi_2 \urcorner), \ldots, \mathit{set}(\ulcorner \varphi_k \urcorner) \ldots))$$

Finally, if T is an $\mathcal{L}_1$ term that stands for a set of $\mathcal{L}_0$ formulae, and f is an $\mathcal{L}_1$ term that stands for an $\mathcal{L}_0$ formula, then we write $\mathit{FACT}_1(T, f)$ to indicate that the formula denoted by f is a member of the set denoted by T. Note that the subscript "1" in the name of the predicate is to give the reader some visual clues as to which language this predicate belongs to: that is, it belongs to $\mathcal{L}_1$. We will also use $\mathit{FACT}_n(\cdots)$ predicates further up the hierarchy. For every statement f appearing in the domain theory Δ_0, we need to include in Δ_1 that f is a $\mathit{FACT}_1(\cdots)$ of Δ_0.

$$\mathit{FACT}_1(\Delta_0, f) \quad \text{for each } f \in \Delta_0$$

The next step is to introduce a predicate $\mathit{PRV}_1(\cdots)$, for provability. This is a binary predicate, taking arguments denoting a set of $\mathcal{L}_0$ formulae and an $\mathcal{L}_0$ formula, with the intended interpretation that $\mathit{PRV}_1(T, f)$ means that the formula denoted by f is provable from the theory denoted by T. To ensure that the predicate behaves as intended, we give axioms in Δ_1 that correspond to provability in $\mathcal{L}_0$. So, for example, this axiomatization will include the following, which capture that any member of T is provable from T, two axioms characterising reduction ad absurdum, i.e., that $\neg\neg f \leftrightarrow f$[2], modus ponens, and that if $f \wedge g$ can be proved from T, then so can f and so can g. (Note: In these axioms, and the remainder of the paper, to make formulae more readable, we will adopt the convention that *free variables are assumed to be universally quantified.*)

$$\begin{array}{l} \mathit{FACT}_1(T, f) \to \mathit{PRV}_1(T, f) \\ \mathit{PRV}_1(T, f) \to \mathit{PRV}_1(T, \mathit{not}(\mathit{not}(f))) \\ \mathit{PRV}_1(T, \mathit{not}(\mathit{not}(f))) \to \mathit{PRV}_1(T, f) \\ \mathit{PRV}_1(T, \mathit{imp}(f, g)) \wedge \mathit{PRV}_1(T, f) \to \mathit{PRV}_1(T, g) \\ \mathit{PRV}_1(T, \mathit{and}(f, g)) \to (\mathit{PRV}_1(T, f) \wedge \mathit{PRV}_1(T, g)) \\ \text{etc.} \end{array}$$

It is straightforward to extend these axioms to give an $\mathcal{L}_1$ axiomatization that characterises $\mathcal{L}_0$ provability: for simplicity, we assume a set of axioms that characterises a *complete* proof system for $\mathcal{L}_0$ (see, e.g., [8, pp.55–62]).

Of course, for different purposes, different types of proof may be appropriate in the characterisation of $\mathit{PRV}_1(\cdots)$. We can *tailor* our notion of $\mathcal{L}_0$ provability by choosing different axioms characterising $\mathit{PRV}_1(\cdots)$. For example, if (for some reason) we wanted a notion of provability that did not include the ability to apply the and-elimination rule, then we would omit the fifth axiom for $\mathit{PRV}_1(\cdots)$ from the list above; if we wanted a constructive, intuitionistic notion of proof, then we would give an axiomatization without the second and third axioms, and so on.

Next, we define the subset relation over sets of $\mathcal{L}_0$ formulae as follows.

$$\begin{array}{l} (T_1 \subseteq T_2) \leftrightarrow \\ \quad \forall f \cdot \mathit{FACT}_1(T_1, f) \to \mathit{FACT}_1(T_2, f) \end{array}$$

We now introduce arguments. We use an $\mathcal{L}_1$ function $\langle \cdots \rangle$ of two arguments, which simply makes a tuple out of these arguments; where a is an $\mathcal{L}_1$ term denoting an argument, we use the projection function $\mathit{conc}(a)$ to extract the conclusion from argument a, and $\mathit{supp}(a)$ to extract the support.

[2] Note that we could collapse these two axioms into one biconditional; the rationale for not doing this will become clear in the following section.

$$\begin{array}{rcl} \mathit{conc}(\langle f, T\rangle) & = & f \\ \mathit{supp}(\langle f, T\rangle) & = & T \end{array}$$

We then say that $\langle f, T\rangle$ is a *prima facie* argument if f is provable from T and T is a subset of Δ_0.

$$\begin{array}{l} \mathit{PF}_1(a) \leftrightarrow \\ \quad (\mathit{PRV}_1(\mathit{supp}(a), \mathit{conc}(a)) \wedge (\mathit{supp}(a) \subseteq \Delta_0)) \end{array}$$

(Recall that Δ_0 here is an $\mathcal{L}_1$ constant which denotes the set of $\mathcal{L}_0$ formula characterising the object level domain of discourse.)

A consistent prima facie argument ($\mathit{CPF}_1(a)$) is one whose support is consistent;

$$\begin{array}{l} \mathit{CPF}_1(a) \leftrightarrow \\ \quad (\mathit{PF}_1(a) \wedge \neg \mathit{PRV}_1(\mathit{supp}(a), \ulcorner \mathbf{false} \urcorner)) \end{array}$$

And an argument is a consistent prima facie argument that is minimal, in the sense that no subset of the support is sufficient to serve as a support for the argument.

$$\begin{array}{l} \mathit{ARG}_1(a) \leftrightarrow \\ \quad (\mathit{CPF}_1(a) \wedge \neg \exists T \cdot (T \subseteq \mathit{supp}(a)) \wedge \mathit{CPF}_1(\langle \mathit{conc}(a), T\rangle)) \end{array}$$

Example 2. *Suppose that Δ_0 is as defined in (Ex_1), above. Then, constructing Δ_1 using the axioms and facts as above, we can conclude the following.*

$$\begin{array}{l} \Delta_1 \models_{\mathcal{L}_1} \mathit{ARG}_1(\langle \ulcorner q \urcorner, \{\ulcorner p \urcorner, \ulcorner p \rightarrow q \urcorner\}\rangle) \\ \Delta_1 \models_{\mathcal{L}_1} \mathit{ARG}_1(\langle \ulcorner \neg p \urcorner, \{\ulcorner t \urcorner, \ulcorner t \rightarrow \neg p \urcorner\}\rangle) \\ \Delta_1 \models_{\mathcal{L}_1} \mathit{ARG}_1(\langle \ulcorner \neg s \urcorner, \{\ulcorner p \urcorner, \ulcorner p \rightarrow q \urcorner, \ulcorner (q \vee r) \rightarrow \neg s \urcorner\}\rangle) \end{array}$$

We now formalise the way that arguments may *attack* one another [7]. In the argumentation literature, "a_1 attacks a_2" is roughly interpreted as meaning "a rational agent that accepts a_1 would have to reject a_2". Unfortunately, there is no consensus on the semantics of attacks, and indeed Dung's abstract argumentation theory completely ignores the issue, simply assuming that one is presented with an attack relation. In logic-based argument, there are two widely used notions of attack: *rebuttal* (where the conclusion of one argument is logically equivalent to the negation of the conclusion of the other) and *undercutting* (where the conclusion of one argument is logically equivalent to the negation of some element of the support): see, e.g., [17]. Since rebuttal is inherently symmetric (in the sense that if a_1 rebuts a_2, then by definition a_2 rebuts a_1), its value in the definition of attack has been questioned [2]. For this reason, we will focus on undercutting as the foundation of attack.

We define a two place $\mathcal{L}_1$ predicate $\mathit{ATTACK}_1(\cdots)$, such that $\mathit{ATTACK}_1(a_1, a_2)$ means that a_1 undercuts a_2, in the sense that the conclusion of a_1 is logically equivalent to the negation of some subset of the support of a_2. The formal definition is as follows.

$$\begin{array}{l} \mathit{ATTACK}_1(a_1, a_2) \leftrightarrow \\ \quad \mathit{ARG}_1(a_1) \wedge \mathit{ARG}_1(a_2) \wedge \\ \quad (\exists f \cdot \mathit{FACT}_1(\mathit{supp}(a_2), f) \wedge \mathit{PRV}_1(\emptyset, \mathit{iff}(\mathit{conc}(a_1), \mathit{not}(f))) \end{array}$$

Example 3. *Suppose that Δ_0 is as defined in Example 1, above, and assume Δ_1 is constructed using the axioms and facts as above. Moreover, let $a_1 = \langle \ulcorner q \urcorner, \ulcorner p \rightarrow q \urcorner\rangle$ and let $a_2 = \langle \ulcorner \neg p \urcorner, \ulcorner t \rightarrow \neg p \urcorner\rangle$. Then $\Delta_1 \models_{\mathcal{L}_1} \mathit{ATTACK}_1(a_2, a_1)$.*

Now, it is well-known that an attack relation is not in itself generally sufficient to resolve the issue of which arguments should be judged acceptable. Considering the various notions of acceptability from [7], for example, preferred extensions and grounded extensions always exist but may be empty, while stable extensions are never empty, but may not exist. More generally, however, Bench-Capon has argued, taking his cue from Perelman [14], that a logical approach is just too simplistic in many scenarios [1, pp.429–430]: to resolve the argument system, we need to consider the values that arguments appeal to, and make our judgements not only the logical soundness of the arguments, but also on how we rank the values embodied in arguments:

> Often, no conclusive demonstration of the rightness of one side is possible: both sides will plead their case, presenting arguments for their view as to what is correct. Their arguments may all be [logically] sound. But their arguments will not have equal value for the judge charged with deciding the case: the case will be decided by the judge preferring one argument over another. [...] One way of [justifying such preferences] is to relate the arguments to the purposes of the law under consideration, or the values that are promoted by deciding for one side against the other.

Bench-Capon goes on to show how Dung's argument framework may be extended with values, intended to capture such a system of arguments: we will proceed to formalise Bench-Capon's framework within $\mathcal{L}_1$. First, we assume that the domain of $\mathcal{L}_1$ contains a set of *values*. We shall not be concerned with the nature of such values, but examples might include, (taking from a legal setting), the right to life, the right to free speech, public interest, and the right to own property. Now, we will associate each argument with such a value, by means of a two-place $\mathcal{L}_1$ predicate $\mathit{VAL}_1(\cdots)$. We require that every possible $\mathcal{L}_1$ argument has a value.

$$\mathit{ARG}_1(a) \rightarrow \exists v \cdot \mathit{VAL}_1(a, v)$$

While one could in principle consider arguments being associated with more than one value, for simplicity we will assume that arguments have exactly *one* value.

$$\mathit{ARG}_1(a) \wedge \mathit{VAL}_1(a, v_1) \wedge \mathit{VAL}_1(a, v_2) \rightarrow (v_1 = v_2)$$

Next, we introduce *audiences*. We assume the domain of $\mathcal{L}_1$ contains a set of audiences: an audience, in Perelman and Bench-Capon's frameworks, is a group of agents who have *preferences* over values. We denote audiences by q, q', and so on, and use a ternary $\mathcal{L}_1$ predicate $v_1 \succ_q v_2$, with the intended meaning that audience q ranks value v_1 above value v_2. The $\succ_a$ relation is assumed to be transitive, irreflexive, and asymmetric, giving the following three axioms for Δ_1.

$$\begin{array}{l} ((v_1 \succ_q v_2) \wedge (v_2 \succ_q v_3)) \rightarrow (v_1 \succ_q v_3) \\ \neg(v_1 \succ_q v_1) \\ (v_1 \succ_q v_2) \rightarrow \neg(v_2 \succ_q v_1) \end{array}$$

We have a ternary $\mathit{DEFEATS}_1(\cdots)$ predicate, with the idea being that $\mathit{DEFEATS}_1(a_1, a_2, q)$ if argument a_1 attacks a_2 and it is not the case that the value promoted by a_1 is ranked over that promoted by a_2 for audience q.

$$\begin{array}{l} \mathit{DEFEATS}_1(a_1, a_2, q) \leftrightarrow \\ \quad \mathit{ATTACK}_1(a_1, a_2) \wedge \\ \quad \mathit{VAL}_1(a_1, v_1) \wedge \mathit{VAL}(a_2, v_2) \rightarrow \neg(v_2 \succ_q v_1) \end{array}$$

We will assume that some appropriate axiomatization is given in $\mathcal{L}_1$ for working with sets of arguments, defining set membership for arguments ("$a \in A$") and subsets ("$A_1 \subseteq A_2$") – the axiomatization is standard, and we thus omit it. We then say an argument a is acceptable with respect to a set of arguments A for audience q if every possible argument that defeats q is itself defeated for q by some member of A [1]. We characterise this via the $\mathcal{L}_1$ predicate $\mathit{ACCEPTABLE}_1(\cdots)$.

$$\begin{array}{l} \mathit{ACCEPTABLE}_1(a_1, A, q) \leftrightarrow \\ \quad \forall a_2 \cdot \mathit{DEFEATS}_1(a_2, a_1, q) \rightarrow \\ \qquad \exists a_3 \cdot (a_3 \in A) \wedge \mathit{DEFEATS}_1(a_3, a_2, q) \end{array}$$

A set of arguments A is conflict free for audience q if for every pair of arguments a_1, a_2, either it is not the case that a_1 defeats a_2, or else a_2 is ranked over a_1 by q.

$$\begin{array}{l} \mathit{CFREE}_1(A, q) \leftrightarrow \\ \quad (a_1 \in A) \wedge (a_2 \in A) \rightarrow \\ \qquad ((\neg \mathit{DEFEATS}_1(a_1, a_2, q)) \vee \\ \qquad (\mathit{VAL}_1(a_1, v_1) \wedge \mathit{VAL}_1(a_2, v_2) \rightarrow (a_2 \succ_q a_1))) \end{array}$$

A set of arguments A that is conflict free for audience q is *admissible* if every argument in the set if acceptable with respect to A.

$$\begin{array}{l} \mathit{ADM}_1(A, q) \leftrightarrow \\ \quad \mathit{CFREE}_1(A, q) \wedge \\ \quad \forall a \cdot (a \in A) \rightarrow \mathit{ACCEPTABLE}_1(a, A, q) \end{array}$$

Finally, a set of arguments A is a *preferred extension* for audience q if it is a maximal (with respect to set inclusion) admissible set with respect to q.

$$\begin{array}{l} \mathit{PE}_1(A_1, q) \leftrightarrow \\ \quad \mathit{ADM}_1(A_1, q) \wedge \forall A_2 \cdot (A_1 \subseteq A_2) \rightarrow \neg \mathit{ADM}_1(A_2, q) \end{array}$$

Thus far, we have shown how a logic based argument system can be developed within our framework that combines such frameworks with Dung's and Bench-Capon's systems. Note that in order to do this, we have frequently defined predicates that take as their argument formulae and sets of formulae: and any mathematically sound framework which achieved this would inherently have to be meta-logical.

3.3 Meta-Arguments: Δ_2

To construct Δ_2, we proceed much as we did when constructing Δ_1. (Note that in this section, many of the definitions are exact analogues of those appearing at level 1, with the predicate subscripts simply changed from 1 to 2: we will omit such definitions when there is no possibility of ambiguity.) First, we will have a constant Δ_1, which will denote the level 1 theory constructed as above (and of course, this level 1 theory was constructed with respect to the level 0 theory Δ_0, containing object-level sentences). We use a quoting convention for formulae in exactly the same way that we used such a convention in Δ_1, and introduce predicates $\mathit{FACT}_2(\cdots)$ and $\mathit{PRV}_2(\cdots)$ and subset relation $\subseteq$ as above. Also analogously to Δ_1, we assert that every statement appearing in Δ_1 is a $\mathit{FACT}_2(\cdots)$ of Δ_1:

$$\mathit{FACT}_2(\Delta_1, f) \quad \text{for each } f \in \Delta_1$$

We also construct a predicate $\mathit{ARG}_2(\cdots)$, which characterises an argument at level 2 of the hierarchy, by way of predicates $\mathit{PF}_2(\cdots)$ (for prima facie level 2 arguments), and $\mathit{CPF}_2(\cdots)$ (for consistent prima facie level 2 arguments), again following the pattern established at level 1.

EXAMPLE 4. *Suppose that Δ_0 is as defined in Example 1, above, and Δ_1 and Δ_2 are constructed as indicated above. Then we can conclude the following.*

$$\begin{array}{l} \Delta_2 \models_{\mathcal{L}_2} \\ \ \exists T.(T \subseteq \Delta_1) \wedge \\ \ \mathit{ARG}_2(\langle \ulcorner \mathit{ARG}_1(\langle \ulcorner q \urcorner, \{\ulcorner p \urcorner, \ulcorner p \rightarrow q \urcorner\}\rangle) \urcorner, T\rangle) \end{array}$$

$$\begin{array}{l} \Delta_2 \models_{\mathcal{L}_2} \\ \ \exists T.(T \subseteq \Delta_1) \wedge \\ \ \mathit{ARG}_2(\langle \ulcorner \mathit{ARG}_1(\langle \ulcorner \neg p \urcorner, \{\ulcorner t \urcorner, \ulcorner t \rightarrow \neg p \urcorner\}\rangle) \urcorner, T\rangle) \end{array}$$

$$\begin{array}{l} \Delta_2 \models_{\mathcal{L}_2} \\ \ \exists T.(T \subseteq \Delta_1) \wedge \\ \ \mathit{ARG}_2(\langle \ulcorner \mathit{ARG}_1(\langle \ulcorner \neg s \urcorner, \{\ulcorner p \urcorner, \ulcorner p \rightarrow q \urcorner, \ulcorner (q \vee r) \rightarrow \neg s \urcorner\}\rangle) \urcorner, T\rangle) \end{array}$$

This example may at first sight not appear to be saying anything more interesting than was said at level 1: indeed, it looks rather like we are saying, in a fancy way, that certain structures may be proved to be level 1 arguments – which we could also say at level 1! To illustrate the value of this construction, let us therefore take apart the reasoning process through which we can assert that a structure is an argument in Δ_2.

Suppose that, for some $\mathcal{L}_0$ formula f and set of $\mathcal{L}_0$ formulae T_1, we have the following:

$$\Delta_2 \models_{\mathcal{L}_2} \mathit{ARG}_2(\langle \ulcorner \mathit{ARG}_1(\langle f, T_1\rangle) \urcorner, T_2\rangle)$$

This is stating that we can prove that in level 2 that $\mathit{ARG}_1(\langle f, T_1\rangle)$: T_1 serves as the support of this argument, and will be a minimal set of $\mathcal{L}_0$ formulae from the domain theory sufficient to establish the $\mathcal{L}_0$ conclusion f.

But what exactly is T_2 here? It is not a set of $\mathcal{L}_0$ formulae, because we are working in Δ_2. T_2 serves as the support for the conclusion $\mathit{ARG}_1(\langle f, T_1\rangle)$: as the subscript indicates, this conclusion is a sentence of $\mathcal{L}_1$, and so the support is a set of $\mathcal{L}_1$ sentences. What will this support look like? That is, what will T_2 contain? It will contain a minimal consistent set of $\mathcal{L}_1$ sentences that are sufficient to establish the conclusion $\mathit{ARG}_1(\langle f, T_1\rangle)$. In particular, T_2 *must contain a minimal set of sentences from Δ_1 that are required to prove that the structure is an argument: in particular, the $\mathcal{L}_1$ axioms corresponding to proof rules that are required to establish this conclusion, and the axioms corresponding to the definition of an argument.*

But of course, this lays bare the mechanism by which we can establish a statement such as $\mathit{ARG}_1(\langle f, T_1\rangle)$: when we are presented with an argument at $\mathcal{L}_2$ to the effect that something is an $\mathcal{L}_1$ argument, we can examine the support to see *how* this conclusion is justified. This justifies our claim, above, that at level 2, we can not only state that a particular structure is an argument, but also we can characterise *the* *means by which we can assert this*, i.e., the mechanism of establishing that something is an argument. This is critical if we want to consider the axioms and rules that were used to construct the argument.

To see the value of this, let us consider, for example, an *intuitionistic* argument to be one that can be established without the use of the axiom $\neg\neg f \rightarrow f$ (cf. [6]). Recall that in our Δ_1 axiomatization, we included an axiom capturing this axiom, which is valid in classical logic, but is not valid in intuitionistic logic. Let RA be an $\mathcal{L}_2$ constant that denotes this $\mathcal{L}_1$ axiom. We can define an $\mathcal{L}_2$ predicate $IARG_2(\cdots)$ which characterises an $\mathcal{L}_2$ argument that can be constructed *without* RA.

$$\begin{aligned}&IARG_2(a) \leftrightarrow\\ &\quad ARG_2(a) \wedge \neg(FACT_2(supp(a), RA))\end{aligned}$$

In general, there will of course be cases where we have $ARG_2(\langle \ulcorner ARG_1(a) \urcorner, T\rangle)$ but not $IARG_2(\langle \ulcorner ARG_1(a) \urcorner, T\rangle)$.

In the same way, we can define conclusions that can *only* be established by means of classical constructions, i.e., cannot be established intuitionistically. Let us define a unary $\mathcal{L}_2$ predicate $SCARG_2(\ldots)$, which takes an $\mathcal{L}_1$ argument, and which is true when this argument can *only* be established classically.

$$\begin{aligned}&SCARG_2(a) \leftrightarrow\\ &\quad \forall T \cdot ((T \subseteq \Delta_1) \wedge ARG_2(\langle \ulcorner ARG_1(a) \urcorner, T\rangle)) \rightarrow\\ &\qquad \neg IARG_2(\langle \ulcorner ARG_1(a) \urcorner, T\rangle)\end{aligned}$$

Again, we note that this type of construction cannot be achieved at lower levels of the hierarchy.

4. CONCLUSIONS

We have argued that the any proper formal treatment of logic-based argumentation must be a *meta-logical* system. This is because formal arguments and dialogues do not just involve asserting the truth or falsity of statements about some domain of discourse: they involve making arguments *about* arguments, and potentially higher-level references (i.e., arguments about arguments about arguments). To illustrate this meta-logic approach to argumentation, and provide a proof of concept for it, we developed a formalisation of argumentation using a hierarchical first-order meta-logic. We defined three tiers of a hierarchical argument system, with the level 0 of this hierarchy corresponding to object-level statements about the domain, level 1 defining the notion of an argument, and capturing notions of attack/defeat, values and audiences, and the acceptability of argument sets. At level 2 of the hierarchy, we are able to reason about the process of asserting that a particular structure represents an argument, and how such an assertion is constructed. In particular, we are able to capture at level 2 the axioms/rules that must be used in order to construct an argument, and hence distinguish between arguments constructed in different ways.

Although meta-*logical* systems have been widely studied in the past four decades, comparatively little research appears to have addressed the issue of meta-*argument*. One notable exception is the work of Brewka, who in his [4], presented a tiered argument system which at first sight appears to have much in common with our own. However, although there are several points of similarity, there are also many differences, and the motivation and ultimate formalisation in Brewka's approach is in fact rather different.

There are several potentially interesting avenues for future work. First, it we believe it would be straightforward to implement such a hierarchical argument system: in particular, PROLOG has been found to be an extremely useful tool for meta-logical reasoning and the implementation of meta-interpreters for logics [21]. Second, our system currently has no notion of dialogue or argumentation protocol: again, it would be straightforward to extend the framework with dialogues, axiomatizing the protocol rules within the system. Third, it would be useful to extend the framework to include reasoning about each agent's beliefs and intentions, as in [12]: as ¡demonstrated in [10, 9], (hierarchical) meta-logic can be an extremely useful tool for this purpose.

Acknowledgments

This work was made possible by funding from NSF #REC-02-19347, NSF #IIS 0329037, and the EC's IST programme under the "ASPIC" project. We gratefully acknowledge the comments of ASPIC researchers Trevor Bench-Capon and Sylvie Doutre, as well as the anonymous reviewers, which have helped us to improve this paper significantly.

5. REFERENCES

[1] T. J. M. Bench-Capon. Persuasion in practical argument using value based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.

[2] P. Besnard and A. Hunter. A logic-based theory of deductive arguments. *Artificial Intelligence*, 128:203–235, 2001.

[3] A. Bondarenko, P. M. Dung, R. A. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93(1-2):63–101, 1997.

[4] G. Brewka. Dynamic argument systems: A formal model of argumentation processes based on situation calculus. *Journal of Logic and Computation*, 11(2):257–282, 2001.

[5] S. Costantini. Meta-reasoning: A survey. In A. C. Kakas and F. Sadri, editors, *Computational Logic: Logic Programming and Beyond – Essays in Honour of Robert A. Kowalski (LNAI Volumes 2408)*, pages 253–288. Springer-Verlag: Berlin, Germany, 2002.

[6] M. Dummett. *Elements of Intuitionism*. Oxford University Press: Oxford, England, 1977.

[7] 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.

[8] M. R. Genesereth and N. Nilsson. *Logical Foundations of Artificial Intelligence*. Morgan Kaufmann Publishers: San Mateo, CA, 1987.

[9] J. Grant, S. Kraus, and D. Perlis. A logic for characterizing multiple bounded agents. *Autonomous Agents and Multi-Agent Systems*, 3(4):351–387, 2000.

[10] K. Konolige. A first-order formalization of knowledge and action for a multi-agent planning system. In J. E. Hayes, D. Michie, and Y. Pao, editors, *Machine Intelligence 10*, pages 41–72. Ellis Horwood: Chichester, England, 1982.

[11] P. Krause, S. Ambler, M. Elvang-Gøransson, and J. Fox. A logic of argumentation for reasoning under uncertainty. *Computational Intelligence*, 11:113–131, 1995.

[12] S. Parsons, C. A. Sierra, and N. R. Jennings. Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, 8(3):261–292, 1998.

[13] S. Parsons, M. Wooldridge, and L. Amgoud. Properties and complexity of some formal inter-agent dialogues. *Journal of Logic and Computation*, 13(3):347–376, 2003.

[14] C. Perelman and L. Olbrechts-Tyteca. *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press: Notre Dame, 1969.

[15] D. Perlis. Languages with self reference I: Foundations. *Artificial Intelligence*, 25:301–322, 1985.

[16] D. Perlis. Meta in logic. In P. Maes and D. Nardi, editors, *Meta-Level Architectures and Reflection*, pages 37–49. Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1988.

[17] J. L. Pollock. Justification and defeat. *Artificial Intelligence*, 67:377–407, 1994.

[18] H. Prakken and G. Vreeswijk. Logics for defeasible argumentation. In D. Gabbay and F. Guenther, editors, *Handbook of Philosophical Logic (second edition)*. Kluwer Academic Publishers: Dordrecht, The Netherlands, 2001.

[19] C. Reed. Dialogue frames in agent communication. In *Proceedings of the Third International Conference on Multi-Agent Systems (ICMAS-98)*, pages 246–253, Paris, France, 1998.

[20] C. Sierra, N. R. Jennings, P. Noriega, and S. Parsons. A framework for argumentation-based negotiation. In M. P. Singh, A. Rao, and M. J. Wooldridge, editors, *Intelligent Agents IV (LNAI Volume 1365)*, pages 177–192. Springer-Verlag: Berlin, Germany, 1998.

[21] L. Sterling and E. Shapiro. *The Art of Prolog (Second Edition)*. The MIT Press: Cambridge, MA, 1994.

[22] S. Toulmin. *The Uses of Argument*. Cambridge University Press: Cambridge, England, 1958.

[23] R. Turner. *Truth and Modality for Knowledge Representation*. Pitman Publishing: London, 1990.

[24] D. N. Walton and E. C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press, Albany, NY, 1995.