Towards an Argumentation-Based Model of Social Interaction

Elizabeth Sklar[1,2], Simon Parsons[1,2], and Munindar P. Singh[3]

[1]Brooklyn College and [2]The Graduate Center
The City University of New York, New York, USA
[3]North Carolina State University
Raleigh, North Carolina, USA
sklar@sci.brooklyn.cuny.edu,
singh@ncsu.edu,
parsons@sci.brooklyn.cuny.edu

Abstract. The application of argumentation to interactive systems is considered, with a focus on social situations in which humans must make complex decisions. Although argumentation has its roots in philosophy, the ideas have been carried into the multiagent systems community and formalized in logic. This rich tradition has provided a solid foundation and has identified a range of proven properties about argumentation as a formal model of interaction between two or more agents, applied to a variety of contexts. Recent work has begun to investigate the application of argumentation in human-agent settings, where there arises a need to model aspects of human interaction that are less structured than agent-only interactions. This need is explored here: a social scenario is presented, issues are identified, and constructs from formal object-level and metalevel argumentation are combined to produce a feasible approach for future implementation in an interactive system designed for human users.

1 Introduction

Recent years have seen formal argumentation develop as a strong intellectual discipline within the realms of artificial intelligence and multiagent systems (MAS). Argumentation is now beginning to be applied in real-world settings, including facilitating public discourse [34] and security [5]. Much of the research in argumentation within the MAS community focuses on premises, arguments, and conclusions—without adequate regard to peculiarly human qualities, such as cognitive biases and limitations. For example, *rationality* in an agent ignores notions of entrenchment, suspension of belief (or disbelief), intuition, defiance, and attitudes towards authority. As a result, the full range of subtleties involved in human social relationships not only has not been modeled, but also cannot be modeled without relaxing some aspects of traditional formalisms or combining aspects of multiple traditions in representating formalisms.

The field of argumentation has historically developed from a study of human-to-human interactions (chiefly in philosophy) to adaptation for agent-to-agent settings (chiefly in computer science). Our concern is that a number of assumptions made in agent-agent settings can weaken the application of argumentation in human-agent scenarios. For example, the assumption of *rationality* is often made in agent-agent settings,

but frequently violated by humans. The rationality principle holds that an agent will choose an action that maximizes its expected utility. That is, a rational agent will avoid an action that costs more than its expected benefit to the agent. However, humans often make decisions based on non-quantifiable considerations, for example, when a person chooses the welfare of another individual either highly valued (such as a spouse or child) or even a stranger over oneself. Though, upon reflection, this behavior can be understood as being rational, it relies upon postulating a possibly complex utility function for the human. In addition, as is well-known, humans exhibit at best a form of bounded rationality [25]. Even in the agent-only setting, the rationality assumption holds only partially since agents are not able to compute the consequences of their beliefs; indeed, the same shortcoming holds for humans [12]. In other words, given a particular utility function, a human may not act so as to maximize his or her utility because of failing to have computed the possible outcomes of the actions under consideration.

Thus, the assumption of rationality is not universally or uniformly preserved when applied to humans in social situations. Therefore, we need to be able to accommodate practical aspects of limited rationality within a framework of reasoning. We propose to address these by extending existing frameworks to incorporate a model of social biases. Moreover, agents may not be able to introspect on their reasoning and would be unable to reliably define, *a priori*, their personal rules for making different choices under every circumstance. Additionally, an agent may be unwilling to admit or be deceptive about such reasoning. As mentioned above, some work in the philosophical literature of argumentation has addressed social biases, in particular the work of Walton, e.g., [33], has been influential. Much less has been done in the area of computational argumentation, though lying [9, 29, 30] (which we do not consider here) has received some attention.

The topics we address here should help to facilitate successful application of MAS-style argumentation in practical human-agent settings. We provide a running example in Section 2. In Section 3, we review some background on argumentation, at both the object and meta levels, in order to clarify the basis of our contribution. Section 4 illustrates the need for incorporating social biases within the formal mechanism. Then, in Section 5, we return to our running example and illustrate how the biases can now be incorporated using our modified system. Finally, we close with a discussion (Section 6) and a summary (Section 7).

2 Running Example

During a national weather emergency, the Princeton-Plainsboro Teaching Hospital in New Jersey loses power and streets nearby begin to flood[1]. Dr Cuddy, the hospital administrator, needs to evacuate the intensive care unit (ICU) patients to a safe place, which could be any of several hospitals in the region that are still functioning. However, doing so requires complex decision making in order to minimize the risk to the health and safety of the patients and hospital staff, as well as limiting liability and expense. The following facts and considerations bear on this situation:

[1] Any resemblance to events, actual or fictional, is purely coincidental [13].

– The National Weather Service reports that a major hurricane is moving slowly up the east coast of the United States and will disrupt normal activity for the next 24 to 48 hours.
– Traveling in the same direction as the hurricane is fraught with danger.
– The New Jersey governor has declared a "state of emergency" and restricted all travel on major roads to emergency vehicles and essential personnel.
– There have been some isolated reports of tornadoes to the west and southwest, in eastern Pennsylvania.
– There has been an unusually strong storm surge along the Delaware coast, and a blogger reports that the storm has knocked out the Salem Nuclear Power Plant located there. Concern arises about the possibility of an incident resembling the Fukushima Daiichi disaster in Japan (March 2011).
– A tanker truck overturned and caught fire on the southbound New Jersey Turnpike near Princeton.
– A reputed local journalist, Seymour Pulitzer, reports talking to a traffic policeman who said that the tanker had a hazard warning label indicating it was carrying a highly flammable substance.

Decision making in this setting involves weighing the evidence provided by the different reports. The facts and considerations, above, could be captured as formal *arguments*, each claiming the truth or falsity of a premise. The arguments may conflict with each other in various ways. The source of information may be significant. Colleagues may express opinions about the evidence. The ultimate decision (as well as any possible intermediate decisions) may impact persons with whom the decision-maker has personal or professional relationships.

As enumerated below, social aspects of the situation influence the decision-making process for Dr Cuddy and encompass a distinct set of *biases* on the part of her sometimes contentious colleagues:

– Dr Cameron says the National Weather Service (NWS) advises against following the hurricane path and the fact that the recommendation comes from the NWS ought to be conclusive in itself. We refer to this type of bias as *arguing from authority.*
– Dr Traub claims that Seymour Pulitzer is a liar and often tricks police officers into giving testimony that he wants to hear. We refer to this type of bias as an *ad hominem attack.*
– Dr Wilson argues that it is *never* safe to follow a hurricane, no matter what. We refer to this type of bias as *epistemic entrenchment.*
– Dr Chase dismisses the nuclear power plant report, telling Dr Cuddy that no one who blogs for a living could be saying anything close to the truth. We refer to this type of bias as *stereotyping.*
– Dr House insists that there is no need to evacuate since the media is just making a big fuss about the hurricane, which will soon pass. We refer to this type of bias as *defiance.*

The remainder of this paper demonstrates the adaptation and combination of methods from formal object-level and metalevel argumentation in order to adequately model the subtleties in this and other similarly complex social situations. We are motivated by

the growing presence of technological aids for human decision making in today's society. One can easily envision an intelligent assistant embedded in *Siri* [26] or *Iris* [14] that could converse with Dr Cuddy to help her weigh the evidence, consider various angles, and ultimately reach a more informed decision.

3 Technical Background

We now outline the essential technical background that we use to demonstrate how the patterns of reasoning described in the previous section can be captured in argumentation. For this purpose we make use of the formal system from [19, 20] and, as indicated below, we have included updates to reflect more recent changes in the argumentation literature.

3.1 Argumentation

We start with a set of agents $Ags = \{Ag_1, \ldots Ag_n\}$. An individual agent $Ag_i \in Ags$ maintains a knowledge base, Σ_i, containing a set of formulae of a propositional language $\mathcal{L}$. Σ_i may be inconsistent. Agent i also maintains the set of its past utterances, called the "commitment store", CS_i. We refer to this as an agent's "public knowledge", since it contains information that is shared with other agents. In contrast, the contents of Σ_i are "private" to Ag_i.

In the description that follows, we use Δ to denote all the information available to an agent. Thus in an interaction between two agents Ag_i and Ag_j, $\Delta_i = \Sigma_i \cup CS_i \cup CS_j$, so the commitment store CS_i can be loosely thought of as a subset of Δ_i consisting of the assertions that have been made public by Ag_i. In some dialogue games, such as those described by Parsons *et al.* [20], anything in CS_i is either in Σ_i or can be derived from it. In other dialogue games, such as those described by Amgoud *et al.* [2], CS_i may contain assertions that cannot be derived from Σ_i. We define an argument as:

Definition 1. *An **argument** A is a pair (S, p) where p is a formula of some language $\mathcal{L}$ with associated inference mechanism $\vdash_{\mathcal{L}}$ and S is a subset of Δ_i such that:*

1. *S is consistent;*
2. *$S \vdash_{\mathcal{L}} p$; and*
3. *S is minimal, so no proper subset of S satisfying both (1) and (2) exists.*

*S is called the **support** of A, written S = Support(A) and p is the **conclusion** of A, written p = Conclusion(A). Thus we talk of p being* supported *by the argument (S, p).*

In other words, an argument is a pair of support and conclusion. The support is a consistent minimal set of formulae from which the conclusion can be derived using some inference mechanism. We write $\mathcal{A}(\Delta_i)$ to denote the set of all arguments which can be made from Δ_i.

In [19, 20], $\mathcal{L}$ was classical logic. However, the resulting argumentation system was shown to have difficulties handling preferences between arguments[2] consistently [3, 4],

[2] See Definition 3, ahead.

and so for this work we define $\mathcal{L}$ differently. We take $\mathcal{L}$ to be a set of atomic propositions $\mathcal{P}$, and associated with $\mathcal{L}$ is a set of defeasible inference rules $R_{\mathcal{L}}$ of the form $p_1 \wedge \ldots \wedge p_n \Rightarrow c$ where $p_1, \ldots p_n, c \in \mathcal{P}$. In the context of some $\Delta_i \subseteq \mathcal{L}$, a rule $p_1 \wedge \ldots \wedge p_n \Rightarrow c$ sanctions the derivation of c when every p_j in the premises of the rule is either a member of Δ_i, or can be derived as the conclusion of another rule in $R_{\mathcal{L}}$.

We distinguish two subsets of Σ_i. One is Γ_i (from [27, 28]), which itself is split into n subsets:

$$\Gamma_i = \Gamma_i^1 \cup \ldots \Gamma_i^n$$

where each Γ_i^j represents agent Ag_i's beliefs about what agent Ag_j believes. Therefore, Γ_i represents all the information that Ag_i has about the agents in Ags.

The other subset of Σ_i that we distinguish is J_i (from [29]). Defining J_i is helpful because sometimes we want to allow an agent to hold a set of *false beliefs* or *justifications* (for holding those false beliefs) which it distinguishes from the assertions it believes to be true. Previously, Sklar *et al.*[29] used J_i to accommodate situations where an agent may justifiably *lie*, that is, construct arguments that are based on false premises[3]. If we define:

$$T_i = \Sigma_i - J_i$$

as the set of beliefs an agent has that it believes to be true, then alie is any q such that $(S', q) \in \mathcal{A}(\Delta_i)$ and there is at least one $s \in S'$ such that $s \in J_i$. (Thus a lie is the conclusion of an argument that is based on at least one premise that the agent believes to be false, an interpretation that aligns with that in [9].) Sklar *et al.*[29] provide a treatment of lying in this sense. Here, we interpret J_i in a broader sense to facilitate other types of reasoning that may require the ability to express similar notions of disbelief or unjustified belief.

In general, since Δ_i may be inconsistent, arguments in $\mathcal{A}(\Delta_i)$ may conflict. We make this idea precise with the notion of *undermining*[4]:

Definition 2. *Let A_1 and A_2 be arguments in $\mathcal{A}(\Delta_i)$. A_1 **undermines** A_2 iff there is some $\neg p \in Support(A_2)$ such that $p \equiv Conclusion(A_1)$.*

In other words, an argument is undermined if and only if there is another argument which has as its conclusion the negation of an element of the support for the first argument. If A_1 undermines A_2, we talk more broadly of A_1 ***attacking*** A_2. (Other systems of argumentation include other forms of attack between arguments—the precise nature of attack does not concern us here other than to point out that the notion of which arguments attack each other can be determined from the conclusions and supports of the arguments.)

It is typical for an agent Ag_i to have different degrees of belief $bel_i(\cdot)$ for the formulae in Δ_i. In this paper, we assume that these belief values, like those in much of the

[3] In [29], agents are allowed to hypothesize convenient falsehoods—propositions that the agent believes to be false—and J_i is constructed to contain falsehoods that can be used in that way. It could, of course, be used in different ways.

[4] Parsons *et al.*[19, 20] call this "undercutting", but we use the term that fits with Prakken's [23] usage, itself derived from the work of John Pollock [22].

uncertainty handling literature,[5] are between 0 and 1. Thus, if there is some argument $A = (S, p)$ and $A \in \mathcal{A}(\Delta_i)$, then we can compute the belief in an argument from the belief in the formulae that support the argument:

$$bel_i(A) = \bigotimes^{bel}(bel_i(s_1), bel(s_2) \ldots bel(s_n))$$

where $S = \{s_1, \ldots, s_n\}$. Often this function is expanded as:

$$bel_i(A) = bel_i(s_1) \otimes^{bel} bel(s_2) \otimes^{bel} \ldots \otimes^{bel} bel(s_n)$$

Where we need to establish the belief in the conclusion p of A, we set $bel_i(p)$ to be $bel_i(A)$. From these values, we can then establish an order over arguments as follows.

Definition 3. *For an agent Ag_i and a set of belief values for arguments $bel_i(\cdot)$, we can define a **preference** order over arguments $\succeq_i^{bel}$ such that $A_1 \succeq_i^{bel} A_2$ iff $bel_i(A_1) \geqslant bel_i(A_2)$. If this is the case, we say that Ag_i* believes A_1 at least as much as A_2.

In addition, we say that $A_1 =_i^{bel} A_2$ iff $A_1 \succeq_i^{bel} A_2$ and $A_2 \succeq_i^{bel} A_1$; and we say that $A_1 \succ_i^{bel} A_2$ iff $A_1 \succeq_i^{bel} A_2$ and $A_2 \not\succeq_i^{bel} A_1$. As with the notion of belief on which these relations are grounded, we use these relations between the conclusions of arguments when they hold for the arguments themselves, so we may talk, for example, of one conclusion being believed at least as much as another.

We can now define the argumentation system we adopt here:

Definition 4. *An **argumentation system** is a pair:*

$$\langle \mathcal{A}(\Delta_i), \mathcal{R} \rangle$$

where $\mathcal{A}(\Delta_i)$ is defined as above, and $\mathcal{R}$ is a binary relation collecting all pairs of arguments A_1 and A_2 such that A_1 attacks A_2. We write this as: $(A_1, A_2) \in \mathcal{R}$.

Note that our argumentation system is a (rather minimal) version of the ASPIC+ system [23]. Since it only uses defeasible rules (in $R_{\mathcal{L}}$) means [17] that its use of preferences does not lead to the problems of consistency with extensions described in [3, 4].

Given an argumentation system, a natural question to ask is which subset of arguments is reasonable to accept given the various attacks between them. The argumentation literature contains a number of notions of ***acceptability*** that one might adopt and a number of approaches to the computation of the set of acceptable arguments from some $\mathcal{A}(\Delta_i)$. One method for computing acceptability is outlined next.

First, we identify an additional relation between arguments, namely ***defeats***, which helps indicate which arguments are acceptable. For example, we might decide that arguments are immune to attacks from arguments that are less believed; i.e., arguments that are more believed *defeat* arguments they attack that are less believed. Formally, we define a relation $Defeats \subseteq \mathcal{R}$, such that if $Defeats(A_1, A_2)$ then A_1 defeats A_2. With this, we can define a new argumentation system:

$$\langle \mathcal{A}(\Delta_i), Defeats \rangle$$

[5] Especially the three most widely used approaches for handling uncertainty—probability theory [15], possibility theory [10], and Dempster-Shafer theory [24].

We can then capture the idea that arguments are immune to attacks from arguments that are less believed by saying:

$$Defeats(A_1, A_2) \text{ iff } (A_1, A_2) \in \mathcal{R} \text{ and } A_1 \succeq_i^{bel} A_2$$

Recently, attention in the argumentation literature has focussed on a *labeling* approach to computing the set of acceptable arguments [7, 8, 31, 32] (nicely summarized by Baroni *et al.*[6]). This approach is attractive because it is simple to describe and to implement, and so we adopt it here. The approach can be described in terms of a ***labeling function*** LF which maps from arguments to a set of labels $\{\texttt{IN}, \texttt{OUT}, \texttt{UNDEC}\}$. We can then write $in(LF)$ to indicate all arguments that are labelled IN by LF, $out(LF)$ to indicate all arguments that are labelled OUT, and $undec(LF)$ to indicate all arguments that are labelled UNDEC. Note that this loose definition does not specify a correspondence between a labeling and a relation over a set of arguments, so we do that next.

We can specify a relation for a labeling using the notion of defeat and another concept, the idea of ***legality***. For a legal labeling LF, an argumentation framework, $\langle \mathcal{A}(\Delta_i), Defeats \rangle$, and an argument $x \in \mathcal{A}(\Delta_i)$:

1. x is legally IN iff x is labelled IN and every $y \in \mathcal{A}(\Delta_i)$ that defeats x is labelled OUT.
2. x is legally OUT iff x is labelled OUT and there is at least one $y \in \mathcal{A}(\Delta_i)$ that defeats x and is labelled IN.
3. x is legally UNDEC iff there is no $y \in \mathcal{A}(\Delta_i)$ that defeats x such that y is labelled IN, and there is at least one $y \in \mathcal{A}(\Delta_i)$ that defeats x such that y is labelled UNDEC.

Note that the UNDEC state occurs when x cannot be labelled IN (because it has at least one defeater that is not OUT), and cannot be labelled OUT (because it has no IN defeater). If an argument is not legally labelled, it is said to be ***illegally*** labelled. More precisely, an argument is illegally labelled l, where $l \in \{\texttt{IN}, \texttt{OUT}, \texttt{UNDEC}\}$, provided it is not legally labelled l.

With the notion of legality tying labelings to $Defeats$ relations, we can identify acceptable sets of arguments through the notions of *admissibility* and *completeness*. An ***admissible*** labeling has no arguments that are illegally IN, and no arguments that are illegally OUT. A ***complete*** labeling is an admissible labeling that, in addition, has no arguments that are illegally UNDEC. Then, given a complete labeling LF, we have:

1. LF is a *grounded* labeling iff there is no complete labeling with a smaller set of IN arguments.
2. LF is a *preferred* labeling iff there is no complete labeling with a larger set of IN arguments.
3. LF is a *stable* labeling if it contains no UNDEC arguments.

If LF is a grounded labeling, then every x labeled IN by LF is in the grounded extension [11]. If LF is a preferred labeling, then every x labeled IN by LF is in the preferred extension. If LF is a stable labeling, then every x labeled IN by LF is in the stable extension. We can then define sets of acceptable arguments:

Definition 5. *For $T \in \{grounded, preferred, stable\}$, the set of **T-acceptable** arguments for an argumentation system*

$$\langle \mathcal{A}(\Delta_i), Defeats \rangle$$

is defined as those arguments in $\mathcal{A}(\Delta_i)$ that are legally labelled `IN` *by a T labeling.*

If there is a T-acceptable argument for a formula p, then the *status* of p is *T-accepted*; while if there is not an T-acceptable argument for p, the status of p is *not T-accepted*.

3.2 Metalevel Argumentation

We consider *metalevel argumentation* as a crucial framework through which to enable aspects of our expanded application of argumentation. In meta-argumentation [16], not only are arguments and the relations between them represented (for reasoning about "objects", as described in Section 3.1), but also arguments are represented for and against the acceptability of the first ("object-level") set of arguments. These arguments about the acceptability of object-level arguments are referred to as *metalevel arguments*. Identifying the acceptable metalevel arguments helps us to identify which arguments are defeated at the object level. This information then feeds into the labeling process at the object level and identifies the acceptable set of object-level arguments.

To enable us to discuss with some precision how metalevel argumentation can be useful, we briefly introduce a formalization of metalevel argumentation. Modgil and Bench-Capon [16] define a *metalevel argumentation framework*[6] as a tuple:

$$\langle \mathcal{A}(\Delta_i), \mathcal{R}, \mathcal{A}_M, \mathcal{R}_M, \mathcal{C}, \mathcal{L}_\mathcal{C}, \mathcal{D} \rangle$$

where $\mathcal{A}(\Delta_i)$ is a set of arguments and $\mathcal{R}$ is an attack relation on object-level arguments as in the previous section, and $\mathcal{A}_M$ and $\mathcal{R}_M$ are sets of arguments and attacks at the metalevel. $\mathcal{C}$ is a set of claims about the arguments in $\mathcal{A}_M$, $\mathcal{L}_\mathcal{C}$ is the language in which the claims are made, and $\mathcal{D}$ is a set of constraints on the attack relation $\mathcal{R}_M$ that are determined by the claims. Essentially, $\mathcal{C}$ is a mapping from $\mathcal{A}$ to statements. For example:

$$\mathcal{C}(\alpha) = \texttt{justified}(x)$$

says that α is a claim that x is justified.

Modgil and Bench-Capon [16] provide an example metalevel argumentation framework that captures Dung's original argumentation system [11]. Here we extend their framework to demonstrate how one might use the degrees of belief defined for the object-level system (introduced above), and how these are used to determine which arguments are acceptable. In this system, $\mathcal{L}_\mathcal{C}$ includes a set of constants and a set of predicates. The set of constants, $\mathcal{K}$, includes $\ulcorner x \urcorner$ for every $x \in \mathcal{A}$ (it is common practice to quote object-level symbols in this manner to make them constants at the metalevel), and the set of predicates is:

$$\{\texttt{justified}, \texttt{defeats}, \texttt{rejected}\}$$

and has a set of well-formed formulae W defined by the following rules:

[6] This is a less general subset of the system presented in [16], but is sufficient for our present purposes.

1. If $x \in \mathcal{K}$, then $x \in W$
2. If $x, y \in W$, then $(x, y) \in W_{\mathcal{R}}$, $W_{\mathcal{R}} \subset W$
3. If $x \in W$ and $x \notin W_{\mathcal{R}}$, then $\texttt{justified}(x) \in W$
4. If $x \in W$ and $x \notin W_{\mathcal{R}}$, then $\texttt{rejected}(x) \in W$
5. If $x, y \in W$ and $x, y \notin W_{\mathcal{R}}$, then $\texttt{defeats}(x, y) \in W$

where $W_{\mathcal{R}}$ is the set of well-formed formulae that relate to attacks. The language $\mathcal{L}_{\mathcal{C}}$ allows us to talk about any of the constants (which will represent arguments in $\mathcal{A}$), attacks between the arguments, whether arguments are justified or rejected, and whether one argument defeats another.

We next need to define the set of metalevel arguments $\mathcal{A}_M$. It is helpful to think of this as the union of three subsets of arguments:

$$\mathcal{A}_M = \mathcal{A}_{MJ} \cup \mathcal{A}_{MR} \cup \mathcal{A}_{MD}$$

where:

– $\mathcal{A}_{MJ}$ contains arguments that are about justified object-level arguments, i.e.:

$$\alpha \in \mathcal{A}_{MJ}, \mathcal{C}(\alpha) = \texttt{justified}(\ulcorner x \urcorner) \text{ iff } x \in \mathcal{A}$$

– $\mathcal{A}_{MR}$ contains arguments that are about rejected object-level arguments, i.e.:

$$\alpha \in \mathcal{A}_{MR}, \mathcal{C}(\alpha) = \texttt{rejected}(\ulcorner x \urcorner) \text{ iff } x \in \mathcal{A}$$

– $\mathcal{A}_{MD}$ contains arguments that are about which object-level arguments defeat others, i.e.:

$$\alpha \in \mathcal{A}_{MD}, \mathcal{C}(\alpha) = \texttt{defeats}(\ulcorner x \urcorner, \ulcorner y \urcorner) \text{ iff } (x, y) \in \mathcal{R} \text{ and } x \succeq_i^{bel} y$$

In other words, we recognize that argument x defeats argument y if there is an attack between x and y at the object level and it is also the case that x is believed at least as much as y.

Finally, we have the set of constraints on claims $\mathcal{D}$, which enforce consistency between the metalevel statements about the object-level arguments. $\mathcal{D}$ contains:

d_1: if $\mathcal{C}(\alpha) = \texttt{defeats}(X, Y)$ and $\mathcal{C}(\beta) = \texttt{justified}(Y)$, then $(\alpha, \beta) \in \mathcal{R}_M$
which says that a claim that X defeats Y attacks the claim that Y is justified.

d_2: if $\mathcal{C}(\alpha) = \texttt{rejected}(X)$ and $\mathcal{C}(\beta) = \texttt{defeats}(X, Y)$, then $(\alpha, \beta) \in \mathcal{R}_M$
which says that a claim that X is rejected attacks the claim that X defeats Y.

d_3: if $\mathcal{C}(\alpha) = \texttt{justified}(X)$ and $\mathcal{C}(\beta) = \texttt{rejected}(X)$, then $(\alpha, \beta) \in \mathcal{R}_M$
which says that a claim that X is justified attacks the claim that X is rejected.

Together, the above constraints define the contents of $\mathcal{R}_M$, the metalevel attack relation. As Modgil and Bench-Capon [16] show, computing the justified arguments in $\mathcal{A}_M$ identifies the justified arguments in $\mathcal{A}$ consistently across the different definitions of extensions.

3.3 Example

To see how all these bits fit together, let us work through a small example. In this example, agent Ag_1 has the following information in its knowledge base $\Delta_1 = \{a, \neg a\}$, and $R_{\mathcal{L}} = \{a \Rightarrow b\}$. From this information, it can construct two arguments:

$$A_1 = (\{\neg a\}, \neg a)$$
$$A_2 = (\{a, a \Rightarrow b\}, b)$$

While $\mathcal{A}(\Delta_1)$ contains additional arguments, for the purposes of this example we only consider these two, and so consider $\mathcal{R}$ as containing just (A_1, A_2). In this example, let us allocate beliefs as:

$$bel_i(a) = 0.6$$
$$bel_i(\neg a) = 0.9$$

These beliefs are not probabilities (because they do not sum to 1). We will combine them, following Amgoud and Cayrol [1], by taking the belief in an argument to be the minimum of the beliefs in the support. In other words, $\otimes^{bel}$ is $\min$. Thus, $A_1 \succeq_i^{bel} A_2$, because $bel_i(A_1) = 0.9$ and $bel_i(A_2) = 0.6$.

Next, we want to establish the set of acceptable arguments, given our object-level argumentation system. We could use the labeling process from Section 3.1, but instead to demonstrate how the metalevel system works, we use it to determine which arguments are acceptable. Our example contains the set of constants, $\mathcal{K}$:

$$\mathcal{K} = \{\ulcorner A_1 \urcorner, \ulcorner A_2 \urcorner\}$$

where each constant is the name of one of the arguments we have defined at the object level. Then the set of well-formed formulae relating to attacks is:

$$W_{\mathcal{R}} = \{(\ulcorner A_1 \urcorner, \ulcorner A_2 \urcorner), (\ulcorner A_2 \urcorner, \ulcorner A_1 \urcorner)\}$$

The above is the set of all possible attack relations between the two arguments. (Below we derive the actual attack relations from information about the object level.) The full set of well-formed formulae at the metalevel is:

$$W = \{\ulcorner A_1 \urcorner, \ulcorner A_2 \urcorner, \texttt{justified}(\ulcorner A_1 \urcorner), \texttt{justified}(\ulcorner A_2 \urcorner), \texttt{rejected}(\ulcorner A_1 \urcorner), \texttt{rejected}(\ulcorner A_2 \urcorner), \texttt{defeats}(\ulcorner A_1 \urcorner, \ulcorner A_2 \urcorner), \texttt{defeats}(\ulcorner A_2 \urcorner, \ulcorner A_1 \urcorner)\}$$

which adds statements about each argument being justified and selected to the set of possible attacks and the set of constants. We then construct $\mathcal{A}_M$ by applying the rules for constructing $\mathcal{A}_M$, in the order in which they are presented above[7], which yields:

$$\{\texttt{justified}(\ulcorner A_1 \urcorner), \texttt{justified}(\ulcorner A_2 \urcorner), \texttt{rejected}(\ulcorner A_1 \urcorner), \texttt{rejected}(\ulcorner A_2 \urcorner), \texttt{defeats}(\ulcorner A_1 \urcorner, \ulcorner A_2 \urcorner)\}$$

with the `defeats` being added since $A_1 \succeq_i^{bel} A_2$. This will produce a set of attacks that can be summarized by the graph in Figure 1, where nodes are arguments, and a directed

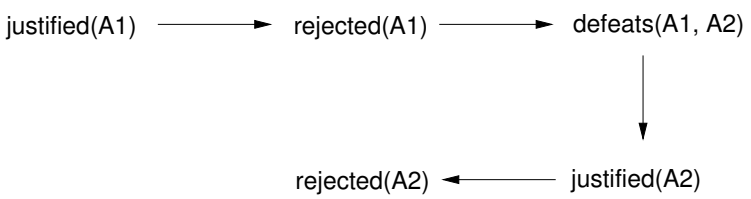

Fig. 1. The attack graph for the example in the case that A_1 defeats A_2

edge between two nodes indicates an attack from the node at the tail of the directed edge to the node at the head of the directed edge.

Running the labeling process from Section 3.1 on this set of arguments with the attack relation of Figure 1[8] will tell us that the only legal labelling has the following arguments (and only the following arguments) labelled IN:

$$\texttt{justified}(\ulcorner A_1 \urcorner), \texttt{defeats}(\ulcorner A_1 \urcorner, \ulcorner A_2 \urcorner), \texttt{rejected}(\ulcorner A_2 \urcorner)$$

and all other arguments are labelled OUT. This resolves the situation at the object level—A_1 is acceptable (in this case grounded-, preferred- and stable-acceptable) and A_2 is not acceptable (that is not grounded-, preferred- or stable-acceptable). The reader can easily verify that this is the same result as one gets by applying the labeling process at the object level—a general proof is provided by Modgil and Bench-Capon [16].

If the beliefs were different, so that $A_1 \not\succeq_i^{bel} A_2$, then $\mathcal{A}_M$ would be:

$$\{\texttt{justified}(\ulcorner A_1 \urcorner), \texttt{justified}(\ulcorner A_2 \urcorner), \texttt{rejected}(\ulcorner A_1 \urcorner), \texttt{rejected}(\ulcorner A_2 \urcorner)\}$$

since there would be no defeats. The attack relation would then include:

$$(\texttt{justified}(\ulcorner A_1 \urcorner), \texttt{rejected}(\ulcorner A_1 \urcorner)), (\texttt{justified}(\ulcorner A_2 \urcorner), \texttt{rejected}(\ulcorner A_2 \urcorner))$$

the attack graph would be as in Figure 2, and we would have:

$$\texttt{justified}(\ulcorner A_1 \urcorner), \texttt{justified}(\ulcorner A_2 \urcorner)$$

as the set of IN arguments. Thus both A_1 and A_2 would be acceptable at the object level. Again, the reader can check that this is the same set of acceptable arguments that would be established by applying the labeling process at the object level.

Since the metalevel approach generates the same set of acceptable arguments as simply computing acceptability at the object level, a reasonable question is why bother with the metalevel? Part of the answer, we suggest, is that the metalevel approach clarifies the computation by explicitly recording the reasons for acceptability (in the first

[7] We note that order of application is important, because changing the order could change the result; but discussion of this is aspect is leaved for follow-on work.

[8] Since the process in Section 3.1 uses a *Defeats* relation between arguments rather than an attack relation, we will need to run the process that we obtain by replacing every instance of the word "defeat" with the word "attack".

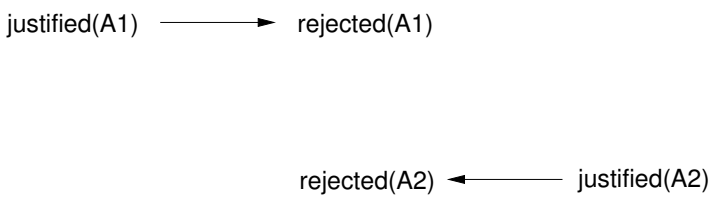

Fig. 2. The attack graph for the example in the case that A_1 does not defeat A_2

part of the example, A_2 is not acceptable because it is defeated by A_1). However, a larger part of the answer is that we can also construct arguments at the metalevel that influence decisions about acceptability, and in the rest of the paper we demonstrate how this ability can be used to capture the situation presented in Section 2.

4 Incorporating Social Biases

Now we return to our opening discussion of social bias and, using the argumentation systems detailed in the previous section, explain how these might be adapted to handle the types of bias highlighted in Section 2. An informal definition of each, within the general context of argumentation, follows.

4.1 Ad Hominem

An *ad hominem attack* is where an agent criticizes an argument because it is provided by a particular agent, rather than based on the content of the argument itself. In the classical interpretation, an *ad hominem* argument is a fallacy. If my roommate tells me that it will rain today and so I should take an umbrella with me when I go out, I may be happy to accept this argument—assuming that I believe my roommate is reliable. However, if my roommate is a notorious practical joker who is always misreporting the weather to trick me into wearing unsuitable clothing, then I might reject the argument exactly because of who it comes from (or, at least, decide to obtain another opinion from a more reliable source).

We can capture an *ad hominem* argument as follows. Imagine we have an argument a at the object level, which is not attacked at the object level. Normally this would mean (as in the second part of the example in Section 3.3) that there were no $defeats(x, \ulcorner a \urcorner)$. An *ad hominem* attack on a could be formalized by:

$$s \in Support(\ulcorner a \urcorner) \wedge s \in CS_j \wedge liar(j) \Rightarrow from\_liar(\ulcorner a \urcorner)$$

and

$$from\_liar(\ulcorner a \urcorner) \Rightarrow defeats(AH, \ulcorner a \urcorner)$$

(where AH indicates "ad hominem") are added to the specification of $\mathcal{A}_M$ so that $defeats(AH, \ulcorner a \urcorner)$ ensures the defeat of a. In other words, if a is based on some fact

from a known liar, then a is defeated by AH. Not only will the object-level system be resolved in such a way that a is not acceptable, but also it will be clear from the metalevel that this is because of the ad hominem argument.

Modeling the argument as explained above makes it possible to distinguish situations in which this ad hominem attack will fail. For example, when my practical joker roommate says it is raining, and I have some independent way to verify that it is raining, then I might decide that the ad hominem attack was not justified in this case. This could be modeled by adding $rejected(AH)$ to the metalevel. This would defeat $defeats(AH, \ulcorner a \urcorner)$ at the metalevel, removing this as a reason for a to be `OUT` at the object level.

4.2 Arguing from Authority

Arguing from authority is where an agent bases its argument on the authority of the agent providing the argument, rather than the evidence supporting it. An argument from authority can be seen as the converse of an *ad hominem* argument, both in the classical interpretation and in the interpretation we offer here. In the classical interpretation, an *ad hominem* argument is an argument against some argument a that can be dismissed. In contrast, an argument from authority is some argument a that can be discounted because it is of the form:

a is true because X said so

and, in the classical view, an argument should be judged on its internal consistency, not by its source. In real life, it can be the case that some arguments from authority hold. In legal cases, expert witnesses are prized exactly because of their ability to make convincing arguments from authority (and they are, of course, subject to *ad hominem* attacks in the courtroom). We can model an argument from authority using this:

$$s \in Support(\ulcorner a \urcorner) \wedge s \in CS_j \wedge authority(j) \Rightarrow from\_authority(\ulcorner a \urcorner)$$

at the metalevel, denoting that a is an argument from authority. Then we can, for example, rule that arguments from authority are more convincing than other arguments that attack them using:

$$from\_authority(\ulcorner a \urcorner) \wedge defeats(\ulcorner b \urcorner, \ulcorner a \urcorner) \Rightarrow defeats(FA, \ulcorner b \urcorner)$$

where FA indicates "from authority". As with the *ad hominem* example, this injects another Defeats relation into the metalevel system, in this case defeating b and so meaning that its attack on a has no effect.

4.3 Epistemic Entrenchment

Epistemic entrenchment is where one agent believes something strongly and nothing will persuade the agent to change its belief. We can represent entrenchment in this notation by saying that agent Ag_i is convinced of conclusion p and nothing that any other agent can say will convince agent Ag_i to change its belief to $\neg p$. In our formal

setting, this relates to an agent having an argument that is based on premises with a higher level of belief than any other information that it holds, or that it can be given. This is a situation akin to that of a conservative politician who asserts that private citizens should be allowed to own guns, and no argument could convince him otherwise. This form of reasoning can be handled using preferences, as in the example from Section 3.3, or we could distinguish the set of unchallengeable beliefs. For example:

$$s \in \mathit{Support}(\ulcorner a \urcorner) \wedge s \in CS_{CD} \Rightarrow \mathit{from\_conservative\_doctrine}(\ulcorner a \urcorner)$$

and

$$\mathit{from\_conservative\_doctrine}(\ulcorner a \urcorner) \wedge \mathit{defeats}(\ulcorner b \urcorner, \ulcorner a \urcorner) \Rightarrow \mathit{defeats}(CD, \ulcorner b \urcorner)$$
$$\mathit{from\_conservative\_doctrine}(\ulcorner a \urcorner) \wedge (\ulcorner a \urcorner, \ulcorner b \urcorner) \in \mathcal{R} \Rightarrow \mathit{defeats}(CD, \ulcorner b \urcorner)$$

will identify any arguments that use facts from the conservative doctrine (CD), will not only prevent them from being defeated by any other argument, but also will ensure that any object-level argument a_1 that is attacked by an argument a_2 which is justified by conservative doctrine, will be defeated by a_2 at the metalevel.

4.4 Stereotyping

Stereotyping is where one participant makes assumptions about the beliefs of another participant, based on the participant's personal attributes such as role or demographics. We can represent stereotyping in this notation by saying that agent Ag_i makes assumptions about the beliefs of agent Ag_j by associating agent Ag_j with a particular *class* of agent. We could[9] therefore copy all of the beliefs of that class into Γ_i^j. Note that this would result in an assumption about agent Ag_i's acceptance of all the beliefs of that particular class, as well as the application of the beliefs of the class to agent Ag_j, as a member of that class. For example, we could say that agent Ag_i believes that all women are in favor of reproductive rights, such as access to birth control and abortion. If agent Ag_j is female, then agent Ag_i will assume that agent Ag_j holds these beliefs and will instantiate these beliefs in Γ_i^j.

The above example is entirely at the object level, but we can also model stereotyping at the metalevel. For example, if we know that the information in Γ_i^j comes from a stereotype, then we can identify this at the metalevel:

$$s \in \mathit{Support}(\ulcorner a \urcorner) \wedge s \in \Gamma_i^j \Rightarrow \mathit{from\_stereotype}(\ulcorner a \urcorner)$$

This can be helpful in cases where the stereotype does not apply. For example, agent Ag_j could hold beliefs contrary to the beliefs of its class, even though it is a member of the class. Alternatively, agent Ag_i could mistakenly associate agent Ag_j with a class when in fact agent Ag_j does not belong to the class. In this case:

$$\begin{aligned} &\mathit{from\_stereotype}(\ulcorner a \urcorner) \wedge \\ &\mathit{associated\_with}(\ulcorner a \urcorner, j) \wedge \\ &\quad \neg \mathit{stereotypical}(j) \Rightarrow \mathit{defeats}(\ulcorner a \urcorner, FS) \end{aligned}$$

[9] There are other ways to handle this, which we will explore in follow-on work.

(where FS means "from stereotype"), capturing the defeat of the argument from stereotype because the person being stereotyped does not belong in the stereotype group.

4.5 Defiance

Defiance is where an agent constructs its arguments from propositions that conflict with its beliefs, on purpose in order to argue in favor of a particular conclusion. We can represent defiance in this notation by saying that agent Ag_i makes arguments that go against its real beliefs. In other words, Ag_i's arguments come from J_i instead of Σ_i. A defiant teenager, for example, will frequently construct *all* her arguments in a dialogue with her parents from J, regardless of what she actually believes—because, as is well documented [35], the developmental state of a teenage brain is such that the teenager often chooses defiant actions over all others, despite the lower expected utility of such choices. Again we can capture this at the object level, but we can also reason about it at the metalevel, for example using:

$$s \in Support(\ulcorner a \urcorner) \wedge s \in J_j \wedge teenager(j) \wedge talking\_to\_parent(j) \Rightarrow from\_defiance(\ulcorner a \urcorner)$$

to identify arguments from defiance. Once this is done, the teenager in question can decide how to deal with them, for example, letting such arguments overrule other arguments:

$$from\_defiance(\ulcorner a \urcorner) \wedge from\_authority(\ulcorner b \urcorner) \wedge defeats(\ulcorner b \urcorner, \ulcorner a \urcorner) \Rightarrow defeats(FA, \ulcorner b \urcorner)$$

(where FA means "from authority") but also allowing specific cases of self-interest to over-rule this over-ruling:

$$from\_defiance(\ulcorner a \urcorner) \wedge from\_authority(\ulcorner b \urcorner) \wedge defeats(\ulcorner b \urcorner, \ulcorner a \urcorner) \wedge something\_I\_desire(\ulcorner b \urcorner) \Rightarrow defeats(SI, FA)$$

(where SI means "self-interest"). This second rule will add $defeats(SI, FA)$ to the metalevel system, defeating the $defeats(FA, \ulcorner b \urcorner)$ generated by the previous rule.

4.6 Combining Biases

The previous sections have sketched how patterns of biased reasoning can be captured in metalevel argumentation, and thus can be used in a rigorous way despite their informal nature. Before returning to our example, we just note that we can use multiple patterns of bias at the same time, since we can use metalevel argumentation to identify how they interact with each other, for example:

$$from\_authority(\ulcorner a \urcorner) \wedge from\_stereotype(\ulcorner b \urcorner) \wedge defeats(\ulcorner b \urcorner, \ulcorner a \urcorner) \Rightarrow defeats(FAS, \ulcorner b \urcorner)$$

where FAS indicates an argument defining the interaction between "from authority" and "stereotype", provides a way to ensure that when an argument from stereotype conflicts with an argument from authority, then the argument from authority wins out.

5 Running Example, Revisited

The biases highlighted in our running example capture how the evidence and opinions interact in combination with each other. Consider Dr Cuddy's decision-making as she weighs the arguments from her colleagues. Note that we ignore the dialogical account of how Dr Cuddy receives the information. Doing so would involve the specification of protocols for exchanging arguments and locutions for passing messages—which could be modelled as by Parsons *et al.* [21]—but would take us far afield from the present topic. Rather we just sketch the arguments that Dr Cuddy might construct and deal with, combining elements of the running example introduced in Section 2 with the extensions to the formal methods described in Sections 3 and 4.

Ad Hominem Attack. We model the argument from Seymour Pulitzer, a renowned purveyor of lies as an example of an ad hominem attack, as follows. Let a be the argument from Pulitzer claiming that a truck with a hazardous substance had overturned on the Turnpike. Let b be the same argument as above (to evacuate patients via the northbound Turnpike). Again, if Dr Cuddy ignored the sources of a and b, then a would attack b based on the priority of preserving safety. However, because $a \in CS_{Pulitzer} \wedge liar(Pulitzer)$, then $defeats(AH, \ulcorner a \urcorner)$; in other words, a would be defeated at the metalevel.

Arguing from Authority. We model the argument from authority, the National Weather Service, as follows. Let a be the argument from the NWS that people should not travel in the path of the hurricane. Let b be an argument from Dr Cuddy to transfer patients to another, better equipped hospital even though it is in the path of the hurricane. If Dr Cuddy ignored the sources of a and b, then b would attack a based on the reasoning that her primary responsibility is to move her patients to safety. However, because $from\_authority(\ulcorner a \urcorner)$, b will be defeated at the metalevel.

Epistemic Entrenchment. Dr Cuddy finds it appropriate to discount entrenchment, so we model the argument from Dr Wilson—one who believes hurricanes should never be followed—as an example of epistemic entrenchment, as follows. Let a be the argument from Dr Wilson (never follow hurricanes). Let b be the same argument as above (to evacuate patients to another hospital located along the path of the hurricane). Dr Wilson's beliefs can be modeled in the same way as those of the conservative politician discussed above, for example by identifying a as being part of the entrenched knowledge, for example $from\_old\_hurricane\_advice(\ulcorner a \urcorner)$, and having such knowledge defeat any arguments, like b, that attack a, adding $defeat(\ulcorner a \urcorner, \ulcorner b \urcorner)$. Dr Cuddy can employ such reasoning to identify what Dr Wilson will say and do, while also being able to compute what arguments are acceptable when this reasoning is rejected and $defeat(\ulcorner a \urcorner, \ulcorner b \urcorner)$ is not considered to hold.

Stereotyping. Dr Cuddy largely discounts a stereotyping argument though she looks for attributes in the stereotype that might be relevant. We model the report from a blogger and Dr Cuddy's derogatory view of bloggers as an example of argument from

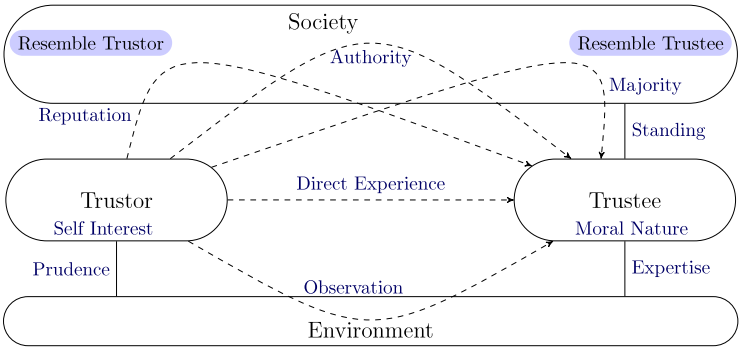

Fig. 3. Relating argumentation schemes to social interaction [18].

stereotype. Let a be the argument from the blogger that the storm may compromise the security of the local nuclear power planted, located nearby in Delaware. Let b be the same argument as above, to evacuate patients to another hospital which happens to be within first-strike range of the power plant. If Dr Cuddy ignored the source of a, then she would accept a, weighing more heavily the risk of exposure to nuclear disaster than the risk of evacuation, and reject b. However, because Dr Cuddy believes that bloggers are unreliable, she will defeat any argument coming from a blogger, meaning that a will be defeated at the metalevel and b will be accepted, since there will no longer be any arguments against it.

Defiance. We model the argument from Dr House that there is no need to evacuate as an example of defiance. Dr Cuddy views Dr House as a defiant colleague who often recklessly ignores authority and the opinions of others. Let a be Dr House's argument against evacuating the patients because he says, in defiance (that is he doesn't believe it but says it just to be defiant) that reports of the severity of the storm are exaggerated. Let b be the same argument as above (to evacuate the patients). If Dr Cuddy were willing to accept Dr House's argument, then she would not evacuate the patients. However, because she knows Dr House to be defiant and hence his argument comes "from defiance", then b will defeat a at the metalevel.

6 Discussion

We motivated this work by considering argumentation-based software agents interacting with people, and we claimed that our work could be used to help those agents better model the people they interact with. Given this motivation, one pertinent question that might be asked about this work is how the biases we have identified and partially formalized would relate to human-agent interactions. We answer this by reference to Figure 3, which comes from our work on formalizing argument schemes for reasoning about trust between agents [18]. Note that these schemes are concerned with reasons

Table 1. Relating forms of social interaction to potential social biases that might arise therein. Here the trustor is the party evaluating an argument and the trustee is the originator of the argument. The biases not discussed in this paper are shown in italics.

Social bias	**Social interaction aspect**
Arguing from authority	Authority; Expertise; Majority; Standing
Defiance	Authority; Expertise; Majority; Standing
Ad hominem attack	Moral nature
Stereotyping	Reputation; Resemble trustee; Resemble trustor
Epistemic entrenchment	Observation; Prudence; Self interest
Over confidence or under confidence	Direct experience
Ad hominem defense	Moral nature

for trust existing between agents. For example A may trust B because it has personal (direct) experience with B.

Figure 3 describes the possible ways in which a trustor may relate to a trustee, and, as we argue in [18], classifies a number of types of social relationship, and interactions that relate to these relationships. These social relationships (including individual traits relevant to social relationships) naturally characterize potential social biases as motivated in this paper, and provide a way to ensure the completeness of a set of social biases. Table 1 shows how the approach described in this paper relates to the above work. It also shows gaps in our current analysis where additional social biases may be identified. (The biases not discussed in this paper are shown in italics.) As Table 1 hints, we can refine our taxonomy of biases by capturing the precise social interaction to which they apply. For example, we might distinguish variants of the arguing from authority bias based on whether it applies to the expertise or social standing of the originator of the argument or whether there is another argument that defends a given argument. We will explore these ideas in future work, as well as exporing other biases, such as the bias introduced by over-confidence in an agent's beliefs when those beliefs turn out to be incorrect.

7 Summary

Argumentation has its roots in work from philosophy that tries to free up reasoning from the limitations of formal logic as a representation of human patterns of reasoning. Computational argumentation has been successful in capturing some of these advances—in particular, the ability to handle inconsistent information—but currently has a much narrower focus than the work in philosophy. Driven by the desire to mechanise reasoning that will be required in human-agent systems, that is, reasoning that acknowledges and models human cognitive biases, this paper has presented a number of patterns of reasoning that we claim are commonly used by human beings and has shown how they can be captured in formal argumentation. To do this, we have suggested using a formal metalevel argumentation system based on [16]. Here we have only sketched the mechanisms that we propose, and our future work will be to develop a more complete account.

Acknowledgments

This research was supported in part by the US Army Research Laboratory through its Network Sciences Collaborative Technology Alliance (NSCTA) under Cooperative Agreement W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Governments is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

The authors wish to thank Sanjay Modgil and Henry Prakken for their help and advice with regard to preference-based argumentation and its implementation within ASPIC+, and Zimi Li for stimulating conversations about object level and metalevel argumentation.

References

1. L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artifical Intelligence*, 34(3):197–215, 2002.
2. L. Amgoud, S. Parsons, and N. Maudet. Arguments, dialogue, and negotiation. In *Proceedings of the Fourteenth European Conference on Artificial Intelligence*, 2000.
3. L. Amgoud and S. Vesic. Repairing preference-based argumentation frameworks. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence*, pages 665–670, 2009.
4. L. Amgoud and S. Vesic. Handling inconsistency with preference-based argumentation. In *Proceedings of the 4th Conference on Scaleable Uncertainty Management*, pages 56–69, 2010.
5. A. Applebaum, K. N. Levitt, J. Rowe, and S. Parsons. Arguing about firewall policy. In *Proceedings of the 4th International Conference on Computational Models of Argument*, pages 91–102, Vienna, Austria, 2012. IOS Press.
6. P. Baroni, M. Caminada, and M. Giacomin. An introduction to argumentation semantics. *The Knowledge Engineering Review*, 2011.
7. M. W. A. Caminada. On the issue of reinstatement in argumentation. In *Proceedings of the 10th European Conference on Logic in Artificial Intelligence*, pages 111–123, Liverpool, UK, 2006.
8. M. W. A. Caminada. An algorithm for computing semi-stable semantics. In *Proceedings of the 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 222–234, Verona, Italy, 2007.
9. M. W. A. Caminada. Truth, lies and bullshit; distinguishing classes of dishonesty. In *Proceedings of the Workshop on Social Simulation*, pages 39–50, 2009.
10. D. Dubois and H. Prade. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York, NY, 1988.
11. 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.
12. J. Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca, 1962.
13. House (TV series), 2012. http://www.usanetwork.com/series/house/.
14. Iris, 2013. https://play.google.com/store/apps/details?id=com.dexetra.iris.

15. E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK, 2003.
16. S. Modgil and T. J. M. Bench-Capon. Metalevel argumentation. *Journal of Logic and Computation*, 6(21):959–1003, 2011.
17. S. Modgil and H. Prakken. A general account of argumentation and preferences. *Artificial Intelligence*, 195:361–397, 2013.
18. S. Parsons, K. Atkinson, K. Haigh, K. Levitt, P. McBurney, J. Rowe, M. P. Singh, and E Sklar. Argument schemes for reasoning about trust. In *Proceedings of the 4th International Conference on Computational Models of Argument*, Vienna, Austria, 2012.
19. S. Parsons, P. McBurney, E. Sklar, and M. Wooldridge. On the relevance of utterances in formal inter-agent dialogues. In *Proceedings of the 6th International Conference on Autonomous Agents and Multi-Agent Systems*, Honolulu, HI, 2007.
20. S. Parsons, M. Wooldridge, and L. Amgoud. On the outcomes of formal inter-agent dialogues. In J. S. Rosenschein, M. Wooldridge, T. Sandholm, and M. Yokoo, editors, *2nd International Conference on Autonomous Agents and Multi-Agent Systems*, New York, NY, 2003. ACM Press.
21. S. Parsons, M. Wooldridge, and L. Amgoud. Properties and complexity of formal inter-agent dialogues. *Journal of Logic and Computation*, 13(3):347–376, 2003.
22. J. Pollock. *Cognitive Carpentry*. MIT Press, Cambridge, MA, 1995.
23. H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1:93–124, 2010.
24. G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.
25. H. A. Simon. *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations*. Free Press, New York, 4th edition, 1997.
26. Siri, 2013. http://www.apple.com/ios/siri/.
27. E. Sklar and M. Q. Azhar. Toward the application of argumentation to interactive learning systems. In *Proceedings of the 8th International Workshop on Argumentation in Multiagent Systems*, Taipei, Taiwan, 2011.
28. E. Sklar and S. Parsons. Towards the application of argumentation-based dialogues for education. In N. R. Jennings, C. Sierra, E. Sonenberg, and M. Tambe, editors, *Proceedings of the 3rd International Conference on Autonomous Agents and Multi-Agent Systems*. IEEE Press, 2004.
29. E. Sklar, S. Parsons, and M. Davies. When is it okay to lie? a simple model of contradiction in agent-based dialogues. In *Proceedings of the First Workshop on Argumentation in Multiagent Systems*, New York , NY, 2004.
30. H. van Ditmarsch, J. van Eijck, F. Sietsma, and Y. Wang. On the logic of lying. In J. van Eijck and R. Verbrugge, editors, *Games, Actions and Social Software*, volume 7010 of *Lecture Notes in Computer Science*. Springer, 2012.
31. B. Verheij. A labeling approach to the computation of credulous acceptance in argumentation. In *Proceedings of the 20th International Joint Conference on Aritificial Intelligence*, pages 623–628, Hyderabad, India, 2007.
32. G. Vreeswijk. An algorithm to compute minimally grounded and admissible defence sets in argument systems. In *Proceedings of the First International Conference on Computational Models of Argument*, pages 109–120, Liverpool, UK, 2006.
33. D. N. Walton. *Ad Hominem Arguments*. University of Alabama Press, 1998.
34. S. Wells and C. Reed. MAgtALO: An agent-based system for persuasive online interaction. In *Proceedings of the AISB Symposium on Persuasive Technology*, pages 29–32, Aberdeen, UK, 2008. The Society for the Study of Artificial Intelligence and Simulation of Behaviour.
35. T. White. Is a happy teenager a healthy teenager?: Four levels of adolescent anger. *Transactional Analysis Journal*, 27(3):192–196, July 1997.