Reaching agreements through argumentation: a logical model (Preliminary Report)*

Sarit Kraus
Dept. of Mathematics
and Computer Science
Bar Ilan Univ.
Ramat Gan, 52900 Israel

Madhura Nirkhe
Dept. of Computer Science
and Institute for Advanced
Computer Studies, Univ. of Maryland,
College Park, MD 20742

Katia Sycara
School of Computer Science
Carnegie Mellon Univ.
Pittsburgh, PA 15213

Abstract

In a multi-agent environment, where self-motivated (selfish) agents try to pursue their own goals, cooperation can not be taken for granted. Cooperation must be planned for and achieved through communication and *negotiation*. We look at argumentation as a mechanism for achieving cooperation and agreement. Using categories identified from human multi-agent negotiation, we present a basis for an axiomatization for argument formulation and evaluation. We offer a logical model of the mental states of the agents based on a representation of their beliefs, desires, intentions and goals. We present argumentation as an iterative process emerging from exchanges among agents to persuade each other and bring about a change in intentions.

1 Introduction

In a multi-agent environment, where self-motivated (selfish) agents try to pursue their own goals, cooperation can not be taken for granted. Cooperation must be planned for and achieved through communication and *negotiation* which inturn often involves *argumentation*.

For example, imagine two mobile robots on Mars, each built to maximize its own utilities. R_1 requests R_2 to dig for a certain mineral. R_2 refuses. R_1 responds with a threat: "If you do not dig for me, I will not help you to transport your samples to the base". R_2 is faced with the task of evaluating this threat. Several considerations must be taken into account, such as whether or not the threat is bounded, what is R_1's credibility, how important is it for R_2 to have help in transporting its samples, so on and so forth. R_1 may take a different approach if R_2 refuses to dig, and respond with "If you dig for me today, I will navigate for you tomorrow." Here, R_2 needs to evaluate the promise of future reward.

Agents may have incomplete kowledge or may lack the ability or the time to make inferences such as in bounded inference systems which either may not be complete or may not be closed under inferences [10, 13]. In order to negotiate effectively, an agent needs the ability to (a) represent and maintain a model of its own beliefs, desires, goals and intentions, (b) reason about other agents' beliefs, desires, goals and intentions, and (c) influence other agents' beliefs, intentions and behavior. Arguments are used by a persuader as a means to dynamically change the preferences and intentions of the persuadee, so as to increase the willingness of the persuadee to cooperate.[1] Over repeated encounters agents may analyze each other's patterns of behavior so as to establish an analog to the human notions of credibility and reputation. This may influence the evaluation of arguments as we will see in scenarios such as "threats" described later. By observing the reactions to the arguments, the sending agent can update and correct its model of the recipient agent, thus refining its planning and argumentation knowledge. Prior work [18]was based on integration of goal graph search, use of multi-attribute utilities, and availability of a case memory of experiences with similar persuadees. In [11] a game theory approach was used for resource allocation with incomplete information, without argumentation, where agents update their beliefs according to the behavior of their opponents.

In this paper we develop a formal logic that forms a basis for the development of a formal axiomatization system for argumentation. We offer a logical model of the mental states of the agents based on a representation of their beliefs, desires, intentions and goals. We present argumentation as an iterative process emerging from exchanges among agents to persuade each other and bring about a change in intentions. Our work on the formal mental model overlaps with others who have developed formal models for communicative agents (e.g., [5, 22, 16, 15, 14])and for mental models of agents (e.g., [21, 19]). The main difference is that we have developed our system from the argumentation point of view. We present a set of axioms that can be used for argument generation and evaluation for automated agents in multi-agent environments.

*This material is based upon work supported in part by the National Science Foundation under Grant No. IRI-9123460. Our paper [8] is an extended version of this paper. e-mails: sarit@umiacs.umd.edu madhura@cs.umd.edu,katia@cs.cmu.edu

[1] Note that we focus only on persuasive arguments, which can be used by automated agents as a mechanism for achieving cooperation. Other argumentations, like argumentation as justification, are beyond the scope of this paper.

2 The Mental Model

We have a set of agents in an environment which is not necessarily cooperative. Their mental states are characterized by using the notions of beliefs, goals, desires, intentions and local preferences. An agent has a set of desires. At any given time, an agent selects a consistent subset of its desires, that serves as its current goals. An agent ascribes different degrees of importance to different goals. It prefers to fulfill goals of higher importance over goals of lesser importance.

The agent's planning process may generate several intentions. Some of these are in what we classify as the "intend-to-do" category and refer to actions that are within the direct control of the agent. Others are among the "intend-that" category [3, 7]. These are propositions that the agent must rely on other agents for satisfying, and are not directly within its realm of control. Often, there is room for argumentation when intend-that actions are part of a plan. Argumentation is the means by which the arguer, the persuader, attempts to modify the intention structure of another agent, the persuadee, to include the intend-that action of the former. While an agent tries to influence the intentions of other agents, other agents may try to convince it as well. Thus, during a negotiation process, an agent may also update its intentions and goals, after receiving a message from another agent.

The agent's belief set includes beliefs concerning the world and beliefs concerning mental states of other agents. An agent may be mistaken in both kinds of beliefs. It may update its beliefs by observing the world, and on receiving messages from other agents. Each agent's actions are based upon its mental model of other agents. The types of arguments (see section 6) that a persuader generates depend on its knowledge of a persuadee's mental model. An important piece of knowledge for argument selecting is a persuader's assessment of the relative importance of the persuadee's goals. For example, a threat is effective if it threatens an important persuadee goal.

3 The Formal Model

We use *minimal structures* [4] style semantics for each of the notions of beliefs, desires, goals, and intentions. The modal operators have certain properties desired from the point of view of our axiomatization. We assume that the agent may not be omniscient (may not have as beliefs all consequences of its "primitive" beliefs [20, 13].) As a result, its set of beliefs may not be consistent and it may not be aware of the inconsistency. As we discuss later, omniscience (or the lack of it) is very important in the context of argumentations. Agents usually negotiate to transfer facts and their conclusions.

The set of desires may not always be consistent either. For example, an agent may desire to earn money today, but also to go on a vacation, and the two desires may lead to a contradiction (see also [19]). Usually, an agent has some preferences among its contradicting desires. The set of goals is a consistent subset of the set of desires. Similarly, we have some implicit properties in mind for actions in the "intend-to-do" category. When an action serves to contribute to one or more of the agent's desires, the agent holds an intention for doing it. The intention may contribute directly to the fulfillment of a desire, or indirectly, through another intention. The action may have a side-effect [5, 2] that does not contribute to any of the agent's desires. In such a case, the agent does not intend the side-effect. Thus we require that the intentions be consistent but not closed under side-effects.[2]

Briefly, we have a set of time lines, each of which extends infinitely far from the past into the future (see [19]). We use time lines instead of more usual worlds because they provide a simple, useful way of incorporating time into our system.

We associate with each time point, time line and predicate, a set of sequences of elements (intuitively, the sequence of elements that have the property of the predicate, at the time point of the time line).

A notion of ***satisfaction*** of a sentence $\psi \in L$ in a time line of a structure, given an interpretation is defined (denoted by $M, l, \bar{v} \models \psi$ see Section 5). The ***truth-set*** of a sentence in the language is the set of time-lines in which the sentence is satisfied, i.e., $||\psi|| = \{l | M, l, \bar{v} \models \psi\}$.[3] A formula is a belief at a given time point at a given time line if its truth-set is belief-accessible. According to this definition, the agent's beliefs are not closed under inferences and it may even believe in contradictions. We will later define different types of agents, in accordance with different properties of their beliefs.

Similarly, we assume that accessibility relations associated with desires, intentions and goals are between time lines and sets of time lines [20]. An agent *intends* (resp.. ***desires*** , has ***goal***) ψ at time t, if the truth-set of ψ ($||\psi||$) is a member of the set of sets of time lines that are intention-accessible (resp.. desires-accessible, goals-accessible) at time t. We further impose restrictions on the set of sets of time-lines that are intention-accessible (resp.. desires-accessible, goals-accessible) to an agent. An agent intends ψ in order to contribute to φ if it intends ψ, intends φ and intends that ψ implies φ. An agent prefers ψ over φ at a give time t, if the agent prefers $||\psi||$ at time t over $||\varphi||$ at time t.

A message may be one of the following types: a request, response or a declaration. A response can be an acceptance or a rejection. A message may carry an argument as a justification. Arguments are produced using special argumentation axioms. An agent can send and receive messages. Unlike Werner's approach [22], we do not assume that receiving a message in itself changes the mental state of the agent. Even receiving an informative message does not change the agent's beliefs, unless it evaluates the message and decides that it should add it

[2] While the issue of how to model the concept of intentions is a very involved topic removed from the main focus of our work, we devote some effort to tailoring our semantics of the intention and desire operators to reflect these desired properties. Our main concern remains in identifying the process of change in these modalities during argumentation.

[3] Note, that if two sentences have the same truth-sets $||\psi|| = ||\varphi||$ then they are ***semantically equivalent***.

to its beliefs[4], especially since we assume that agents are untrustworthy, and may even be untruthful. Only an evaluation process following an argument may change the internal state of the agent.

4 Syntax

We denote by *Agents* the set of agents. We assume that there are four modal operators for agent i: Bel_i for beliefs, $Desire_i$, for desires, $Goal_i$ for goals and Int_i for intentions. It may be the case that the agent is motivated by the need to satisfy its own goals or desires, or that it is convinced to perform an action following an argument.[5] In addition, we assume that there is another modal operator, $Pref_i$ which is associated with the agent's preferences among goals, desires and intentions. Following [19], the basis for our formalism is a simple temporal language. Informally, we have a set of time lines (which play the role of "worlds" in the modal logic). The set of time lines is infinite. We also have a set of time points. At every time in each time line, some propositions are true (and the rest are false).[6]

We have a set TC of time point constants, a set TV of time point variables, a set AC of agent constants, a set AV of agent variables, and a set *Pred* of predicate symbols. We denote by *Variables* the set of all variables (including AV and TV), by *Constants* the set of all constants (including AC and TC), and by *Terms* the set of variables and constants. We also use the symbol *nil*. We first define the set of the formulas of our language.

1. If $t_1, t_2 \in TC \cup TV$ then $t_1 < t_2$ is a wff.
2. If $x_1, x_2 \in Terms$ then $x_1 = x_2$ is a wff.
3. If $P \in Pred$ is a k-ary predicate and $x_1, ..., x_n$ are terms, $t \in TC \cup TV$ and $i \in AC \cup AV$ then , $[t, P(x_1, ..., x_n)]$ is a wff (read as: $P(x_1, ..., x_n)$ *is true at time* t), and so is $[t, Do(i, P(x_1, ..., x_n))]$ is a wff (read as:$P(x_1, ..., x_n)$ *is done by* i *at time* t).
4. If φ is a wff and ψ is a wff then so are $\varphi \wedge \psi$ and $\neg\varphi$. If φ is a wff and $x \in Variables$ then $\forall x \varphi$ is a wff. $\exists, \vee, \rightarrow$ have their usual meanings.
5. If φ and ψ are wffs, $t \in TC \cup TV$, $i, j \in AC \cup AV$ then the following expressions are wffs: $[t, Bel_i \varphi]$ (*i believes φ at time t*), $[t, Desire_i \varphi]$ (*i desires φ at time t*), $[t, Goal_i \varphi]$ (*i has a goal φ at time t*), $[t, Int_i \psi)]$ (*i intends φ*), $[t, Int_i(\varphi, \psi)]$ (*i intends φ at time t to contribute to ψ*), $[t, Pref_i(\varphi, \psi)]$ (*i prefers φ over ψ at time t*) and $Agent(\psi, i)$ (*i is ψ's agent*).
6. If φ and ψ are wffs then $Request(\psi, \varphi)$ (*ψ is requested with the argument φ*), $Reject(\psi, \varphi)$ (*ψ is rejected with the argument φ*), $Accept(\psi, \varphi)$ (*ψ is accepted with the argument φ*), $Decl(\psi)$ (*ψ is declared*), $Accept(\psi)$, $Request(\psi)$, $Reject(\psi)$ are messages.
7. If m is a message, $t \in TC \cup TV$ and $i, j \in AC \cup AV$, then $[t, Receive_{ij} m]$ (*i receives m from j at time t*) and $[t, Send_{ij} m]$ (*i sends m to j at time t*) are wffs.

We will sometimes use the abbreviation $[t, \varphi \wedge \psi]$ for $[t, \varphi] \wedge [t, \psi]$ and will freely interchange $[t, \neg\varphi]$ and $\neg[t, \varphi]$. We will use similar abbreviations for $\vee$ and $\rightarrow$.

5 Semantics

We start with the semantics of the various formulas of our language. This will be followed by the semantics for our modal operators.

Time is a pair $\langle T, \prec \rangle$ where T is a set of time points and $\prec$ is a total order on T (unbounded in both directions).

A model M is a structure $< \Xi, L, Agents, A, B, G, D, I, P, RECEIVE, SEND, \Phi, v, \mathcal{M} >$, where

1. Ξ is a set of elements in the agent's environment and $\mathcal{M}$ is a set of messages
2. L is a set of time-lines
3. *Agents* is a set of agents
4. $B : L \times T \times Agents \rightarrow 2^{2^L}$ is the belief accessibility relation
5. $G : L \times T \times Agents \rightarrow 2^{2^L}$ is the goals accessibility relation
6. $It : L \times T \times Agents \rightarrow 2^{2^L}$ is the intention accessibility relation
7. $D : L \times T \times Agents \rightarrow 2^{2^L}$ is the desire accessibility relation
8. $P : L \times T \times Agents \rightarrow \{< U, U' > | U, U' \in 2^L\}$ is the preference relation
9. Φ interprets predicates and v interprets constants
10. $RECEIVE : L \times T \times Agents \times Agents \rightarrow \mathcal{M}$ indicates the messages received by the agents, $SEND : L \times T \times Agents \times Agents \rightarrow \mathcal{M}$ indicates the messages sent by the agents[7]
11. $A : Pred \times L \times T \rightarrow Agents \cup \{nil\}$ allocates an agent (if any) that performs an action in a given time period. For space reasons, we can't define satisfiability here.

5.1 Properties of the Modalities

In all the following axioms we will assume that the unbounded variables are universally quantified as follows: $\forall l \in L, a \in Agents, \tau, \tau' \in T$. In addition, in all the axiom schemas, we assume that $i \in AC \cup AV$, $t \in TC \cup TV$ and ψ, ϕ and φ can be replaced by any wff in the language.

Let us start with the semantics for the intention operator (It). Following Bratman [2], we would like the

[4] The new information may be inconsistent with the agent's current beliefs. We leave this for future discussion. See for example, [12, 6].

[5] Our intention model is closer to Shoham's *Dec* and [19]'s *Comit* than to Cohen's and Levesque's "Intend".

[6] We have extended [19] to deal with the FOL case. We prefer this approach, where time can be expressed explicitly, over others where time periods cannot be expressed in the language, (for example [5]) since threats and arguments both evolve in time. We use an extension of first order logic, rather an extension of propositional logic since it is useful in the formalism of argumentation.

[7] We note that if for all $l \in L, t \in T$ and $i, j \in Agents$, $RECEIVE(l, t, i, j) = SEND(l, t, j, i)$ then the communication is reliable.

intentions to be consistent. This can be achieved by introducing two constraints on the intention accessibility relation It.

[(CINT1)] $\emptyset \not\in It(l, \tau, a)$.
[(CINT2)] If $U \in It(l, \tau, a)$ and
$V \in It(l, \tau, a)$ then $U \cap V \in It(l, \tau, a)$.

The following axioms (schemas) are sound with respect to the above conditions[8]:

[(INT1)] $[t, \neg Int_i \text{false}]$
[(INT2)] $[t, Int_i \psi] \wedge [t, Int_i \varphi] \rightarrow [t, Int_i \psi \wedge \varphi]$

is a basic premise for the argumentation system. To see this, suppose an agent wants its opponent to intend-to do α that contributes to its intentions and goals. This intention (α) may contradict other intentions of its opponent. Due to the consistency requirement, the agent must convince its opponent (using argumentation) to give up its original contradictory intentions to make place for α.

We would also like the agent's intentions to be closed under consequences, which in turn means that we would like the following axiom to be sound:

[(INT3)] $[t, Int_i \psi] \wedge [t, Int_i \psi \rightarrow \varphi] \rightarrow [t, Int_i \varphi]$

Closure under consequence warrants another restriction on It:

[(CINT3)] If $U \in It(l, \tau, a)$, and $U \subseteq V$, then $V \in It(l, \tau, a)$.

As a result of CINT3 we get the following axiom schema as well:

[(INT4)] $[t, Int_i(\phi \wedge \psi)] \rightarrow [t, Int_i\ \phi] \wedge [t, Int_i \psi]$

The agent does not necessarily intend the side-effects of its intentions. This means that we do not have an axiom of the following form:

$[t, Int_i \phi] \wedge [t, Bel_i \phi \rightarrow \psi] \rightarrow [t, Int_i \psi]$. [9]

We will make similar restrictions on G and obtain similar properties for goals. Intentions and goals are consistent since the agent needs to act and plan according to its intentions and goals. Desires may be inconsistent, however, we don't want the agent to desire false.[10] Usually, an agent has some preferences among its contradicting desires.

We impose the following restrictions on the desires (D) operator:

[(CD1)] $\emptyset \not\in D(l, \tau, a)$.
[(CD2)] If $U \in D(l, \tau, a)$ and $U \subseteq V$ then $V \in D(l, \tau, a)$.[11]

This restrictions yield axioms schema similar to INT1 and INT4 where Int_i is replaced by $Desire_i$.

5.2 Agent types

Within the general framework defined above, it is possible to define various types of agents. We define here some additional conditions on the models that these agents must satisfy in order to have a particular character. In Section 5.4 we discuss how agent types may be guiding factors in the selection of argument categories and generation of arguments.

An Omniscient Agent

An agent whose beliefs are closed under inferences is said to be omniscient. For omniscience we impose the following additional conditions on the model, which render it equivalent to a Kripke Structure.

[(CB1)] $L \in B(l, \tau, a)$.
[(CB2)] If $U \in B(l, \tau, a)$ and $U \subseteq V$ then $V \in B(l, \tau, a)$.
[(CB3)] If $U \in B(l, \tau, a)$ and $V \in (l, \tau, a)$ then $U \cap V \in B(l, \tau, a)$.

The omniscient agent then has the following axioms:

[(B1)] $[t, Bel_i \text{true}]$
[(B2)] $[t, Bel_i \psi \wedge \varphi] \rightarrow [t, Bel_i \psi] \wedge [t, Bel_i \varphi]$
[(B3)] $[t, Bel_i \psi] \wedge [t, Bel_i \varphi] \rightarrow [t, Bel_i \psi \wedge \varphi]$

To ensure that an agent does not believe in false, we need to impose the following restriction:

[(CB4)] $\emptyset \not\in B(l, \tau, a)$.

While all agents are not omniscient, every agent has its intentions and goals closed under consequence. This is justifiable since the set of intentions is much smaller than the set of beliefs. The agent is aware of its intentions since it needs to search for vplans to achieve them. Therefore, they are under its scope and it is reasonable to assume that the agent can compute their closure under consequence.

A Knowledgeable Agent

There are some agents that are knowledgeable, i.e., their beliefs are correct. The corresponding axiom schema is:

[(B5)] $[t, (Bel_i \phi) \rightarrow \phi]$.

The related condition, which makes this axiom sound is the following:

[(B5)] $U \in B(l, \tau, a)$ then $l \in U$.

5.3 Properties Associated with Change in Modalities Over Time

So far, we have considered only local properties. We would like now to consider how the agent's beliefs change over time.

An Unforgetful Agent

An agent who does not forget anything, i.e. one who always has memory of its complete past can be characterized by:

[(BUF1]) If $\tau \prec \tau'$ then $B(l, \tau, a) \subseteq B(l, \tau', a)$

A Memoryless Agent

We would like to capture agents that don't have memory and can't reason about past events. We consider time lines that are restricted to be finite from one side and infinite from the other (i.e., a ray). Let us denote by $l(0)$ the first time period of l. An agent doesn't have a memory under the following condition: If $U \in B(l, \tau, a)$ and $l' \in U$ then $l'(0) = \tau$. Given a time line, the truth value of a sentence in which $[t, Bel_i \psi]$, where

[8] The proof that the a model validates the axioms iff it satisfies the conditions can be found in [20] in the context of propositional epistemic structures.

[9] Although it may be the case that the agent is aware of some implications of its intentions, only if it intends the implication (as dictated by Axiom INT3) does it intend the consequence.

[10] Note that there is a difference between $[t, Desire_i p] \wedge [t, Desire_i \neg p]$ and $[t, Desire_i p \wedge \neg p]$. We allow the first case, but not the second one.

[11] CD1 is similar to CINT1 and CD2 is similar to CINT3

$t' \in TC \cup TV$ appears in ψ and $\bar{v}(t') \preceq \bar{v}(t)$ is not well defined in this case.

Cooperative Agents

A group A of agents $A \subseteq Agents$ is cooperative[12] if they share some common goals. This imposes the following condition: $\bigcap_{j \in A} G(l, t, \bar{v}(j)) \neq \emptyset$.

Furthermore, we require that the goals are common belief.[13] That is, let Δ be the set of common goals $= \bigcap_{j \in A} G(l, t, \bar{v}(j))$ then $[t, C \bigwedge_{\psi \in \Delta} \bigwedge_{j \in A} Goal_j \psi]$.

Our definition of cooperativeness of agents may be time-dependent. A set of agents that are cooperative in a given time period, may become adversaries in later time period, when their common goals do not exist anymore.

5.4 Inter-relations Among Modalities

We have so far presented axioms and semantics conditions to capture properties of each modality. We now investigate inter-relations among the different modalities.

To start with, every goal is also a desire.

[(GD)] $[t, Goal_i(\phi)] \rightarrow [t, Desire_i \phi]$.

An agent adopts all its goals as intentions:

[(GINT)] $[t, Goal_i \phi] \rightarrow [t, Int_i \phi]$.

However, there may be intentions that are not goals. An agent may hold an intention in response to a threat or promise for a reward. During the argumentation the agent may come to have an intention to prevent the opponent from carrying out the threat, or to convince it to give a reward, which only indirectly contributes to one of the agent's goals.

We assume that an agent's intention doesn't contradict its beliefs:

[(INTB)] $[t, Int_i \phi] \rightarrow [t, \neg Bel_i \neg \phi]$.

The corresponding restriction on the It and B relations is as follows:

[(CINTB)] If $U \in It(l, \tau, a)$ then $L - U \notin B(l, \tau, a)$.

We may characterize an agent as **confident** if it believes that it will succeed in carrying out its actions[14]:

[(Conf)] $[t, Int_i \phi] \wedge Agent(\phi, i) \rightarrow [t, Bel_i \phi]$

An agent that is sure that it will be able to satisfy all its intentions, including the ones that are not under its direct control can be said to be overconfident.

[(OverConf)] $[t, Int_i \phi] \rightarrow [t, Bel_i \phi]$

We take a different approach concerning preferences and desires than [21]. We assume that the agent's preferences are over the sets of time lines, while [21]'s preferences are over single models. An agent prefers ϕ over ψ if it prefers the truth-set $(||\phi||)$ over $||\psi||$. In different models, different restrictions may be put on P. For example, P can be transitive.

The agent's desires are not derived from its preferences (see also [9]), but we make the following restriction on the model:

[(CPD)] $\forall U, U' \in 2^L,\ if U \in D(l, a, \tau)$ and $< U', U > \in P(l, \tau, a)$ then $U' \in D(l, \tau, a)$.

Hence, in our model the following axiom is sound:

[(PD)] $[t, Desire_i \psi] \wedge [t, Pref_i(\varphi, \psi)] \rightarrow [t, Desire_i \varphi]$

6 Axioms for Argumentation and for Argument Evaluation

Arguments serve to either add an intention to the persuadee's set or to retract an intention. In each category we present examples that are borrowed from human argumentation as well as examples of automated agents interactions. We have currently identified five types of arguments[15] that are suitable for our framework:

(1) Threats to produce goal adoption or goal abandonment on the part of the persuadee.

(2) Enticing the persuadee with a promise of a future reward.

(3) Appeal to precedents as counterexamples to convey to the persuadee a contradiction between what she/he says and past actions.

(4) Appeal to "prevailing practice" to convey to the persuadee that the proposed action will further his/her goals since it has furthered others' goals in the past.

(5) Appeal to self-interest to convince a persuadee that taking this action will enable achievement of a high-importance goal.

Examples of argumentation among automated agents are based on the scenario described below. Agents with different spheres of expertise may need to negotiate with each other for the sake of requesting each others' services. Their expertise is also their bargaining power. As an example, consider a robot R_e who has a better "eye" (has a powerful camera) while R_h is a better "hand" (has skilled functions with dexterous fingers enabling it to isolate mineral chunks). Yet another agent R_m has specialized maps and terrain knowledge and is adroit at navigation. Imagine these three self-motivated robots with goals to obtain samples from Mars. R_e, R_h and R_m are looking for different mineral samples. We can imagine these three agents facing the need to argue with each other. After the description of each category of arguments, we present an example of its usage in this environment.[16]

Arguments involving threats

Suppose agent j intends that agent i should do α at time $\bar{t}$ and i refuses. Based on its own beliefs, j assumes that i refused to do α probably since α contradicts one of i's goals or intentions. If there is an action β that j can perform, that contradicts (as per j's beliefs) another goal

[12] Among cooperative agents, as among noncollaborative agents, conflicts may occur and negotiation may be required. However, the argumentation is of a different nature, and we shall not dwell on that here.

[13] We denote by $E\psi$ the property that all the agents of A believe ψ. ψ is common knowledge at time t if everyone in A believes ψ, everyone believes that everyone believes ψ and so on. $C\psi$ denotes that ψ is common knowledge.

[14] Note, that the side-effects of its intentions will influence the intentions of a confident agent. If $[t, Int_i \phi] \wedge Agent(\phi, i)$ then i believes that ϕ. Therefore, if it is aware of some side-effects (e.g., $[t, Bel_i \phi \rightarrow \psi]$), then it can't intend the negation of the side-effects (e.g.,$[t, \neg Int_i \neg \psi]$).

[15] These are a subset of those identified in [17]. See also [1].

[16] Similar needs for argumentation can be envisaged when automated agents seek different goals at the site of a nuclear accident, or at the site of a forest fire.

of i, and this last goal is preferred by i (again according to j's beliefs) over the first one, j threatens i that it will do β if i won't do α. This type of argument may appear in several different forms. For example, suppose agent j intends that agent i shouldn't do α , at time $\bar{t}$, and i insists on continuing to intend α. Here, agent j threatens i to do β if i will *do* α.

A labor union insists on a wage increase. The management says it cannot afford it, and asks the union to withdraw its request. The management threatens that if it grants this increase it will have to lay off employees to compensate for the higher operational cost that the increase will entail. The outcome (i.e. whether the union succumbs to the threat or not) depends on the union's preferences. If preserving employment is more important than wage increase, the union will accept the argument (assuming it believes that the management will carry out the threat). If wage increase is more important, then the union will not accept the argument and insist on wage increase (here, whether or not it believes the management will carry out its threat is irrelevant the union's decision.)

One of the questions related to generating a threat is: how does j choose β? If j wants the threat to be effective, carrying out β should be painful for i and conflict one of its goals or intentions (as we stated above). However, the threat should be credible according to i's beliefs (see our discussion concerning the evaluation of threats below). First of all, doing β should be within the power of j (at least in i's view). Furthermore, usually, carrying out a threat may contradict some of j's vintentions or goals. These intentions and goals should be less preferred by j than the goal that α contributes to (again, at least in i's view). There may be several such βs that j may choose from. The β that is chosen depends on whether the persuader, j wants to inflict a very strong threat (i.e., a β which contradicts a preferred goal or intention of i), or to start with a weaker threat (i.e., one which will contradict a less preferred goal of i) and, if i refuses it, escalate with stronger threats (wearing i down).

The first axiom of Figure 1 demonstrates one possibility for a formalism of a creation of a threat argument in the logic presented in Sections 5.

In the example of the robots on Mars; agent R_h must explore in a dimly lit area while digging for its mineral. Some help from R_e in scanning the area with a high resolution camera would contribute heavily towards this goal. R_h requests from R_e the use of its camera. R_e refuses, since the time spent in furthering R_h's goals will interfere with its own goal to dig for its own mineral. R_h then threatens R_e that it will not cooperate in the operation to transport all the collected samples to the spaceship if R_e does not oblige.

Argument evaluation is an important aspect of argumentation. Here, we demonstrate some factors affecting the evaluation of a threat. Since we don't assume that agents are honest, the main problem in the evaluation is how to figure out whether the threatening agent will carry out its threat. Usually, executing a threat will affect the agent that threatens to carry it out, and this has a bearing on the evaluation.

Evaluation of Threats

Suppose j had requested i to do α at a given time point $\bar{t}$, and it had threatened i that if it does not do α, j would do β. Now, i should consider several issues. First of all, how bad is the threat? If α contradicts one of i's goals and β contradicts another goal g, which goal does i prefer? But then again, i should evaluate whether j will at all carry out its threat. We may assume, that β has some negative side-effects to j as well. The question is whether j prefers the benefit for itself from α over the losses from the side-effects of β in case it will carry out the threat. Another issue that is relevant here is how important is it for j to preserve its credibility and reputation. Another issue for i to consider is whether the threat is a *bounded* threat. A bounded threat is always credible since i is aware of prior arrangements made by j to execute the threat should i default. Usually, j will convey this information to i in a prior exchange. If i believes that j may carry out its threat and decides that it is worthwhile for it to do α, it still needs to update its goals and intentions. Here i will intend α in order to contribute to preventing j from doing β which contributes to g. Note, that here i intends α without β being a goal. Furthermore, since any goal is also an intention (GINT), i should abandon the goal that α contradicts, as well as the related intentions.

In the second axiom of Figure 1 (A2), we have listed one way to evaluate whether a threat is credible. Here, i believes that carrying out the threat β by j will contradict i's goal (g_3^i) as well some possible goals of j (g_4^j). If i believes that g_4^j is one of j's goals, and if it believes that j prefers the goal that α contributes to over g_4^j then it will believe that β is a credible threat.

Promise of a future reward

Agent j entices agent i to do action α (alternately, avoid doing α) at time t by offering as a reward, to do an action β at a future time. Agent j believes β to contribute to the desires of i.

An example is of a sales agent trying to persuade customer to buy a VCR by offering a free servicing plan and a set of blank cassettes. For space reasons, we don't give the formal description of the rest of the axioms.

Consider the scenario described in the threat example above involving robots. Instead of responding with a threat, R_h could offer to contribute towards R_e's goal by helping it to isolate its samples better from the debris by using its skills for sorting with its skilled fingers. This would reduce the weight of the samples that R_e now plans to collect, and increase the ratio of mineral to debris greatly for R_e.

Counter Example

Here, agent j had intended that i do α at time $\bar{t}$, requested it from i, but i refused. Now j believes that the reason i refused is that it contradicts one of its goals or intentions. However, j believes, that in the past, i had done another action β that also did contradict the same goal or similar intention, and brings it up as a counterexample.

As an example, consider a parent trying to persuade a teenager to stay up until midnight to study for an exam.

$$\begin{array}{llllll}
\mathbf{A1}: & t_1 < t_2 < t_3 < \bar{t} < t_4 \leq t_5 \wedge i \neq j & \wedge & [t_1, Send_{ji}Request([\bar{t}, Do(i,\alpha)])] & \wedge & [t_2, Receive_{ji}Reject([\bar{t}, Do(i,\alpha)]))\\
\wedge & [t_3, Bel_j([t_3, Goal_i[\bar{t}, g_1] \wedge Goal_i[t_5, g_2]])] & \wedge & [t_3, Bel_j([t_3, Pref([t_5, g_2], [\bar{t}, g_1])])] & \wedge & [t_3, Bel_j[\bar{t}, \alpha \rightarrow \neg g_1]]\\
\wedge & [t_3, Bel_j[t_4, \beta] \rightarrow [t_5, \neg g_2]] & \wedge & [t_3, Bel_j[t_3, Credible(\beta, \alpha, i,)]] & \wedge & [t_3, Bel_j[t_3, Appropriate(\beta, \alpha, i)]]\\
\end{array}$$
$$\rightarrow Send_{ji}Request([\bar{t}, Do(i,\alpha)], \neg[\bar{t}, Do(i,\alpha)] \rightarrow [t_4, Do(j,\beta)])$$

$$\begin{array}{llll}
\mathbf{A2}: & t_1 < t_2 < \bar{t} < t_3 \wedge i \neq j & & \\
\wedge & & [t_1, Receive_{ji}Request([\bar{t}, Do(i,\alpha)], \neg[\bar{t}, Do(i,\alpha)] \rightarrow [t_3, Do(j,\beta)])] & \\
\wedge & [t_2, Bel_i[\bar{t}, \alpha \rightarrow (\neg g_1^i \wedge g_2^j)]] & \wedge & [t_2, Bel_i[\bar{t}, \beta \rightarrow (\neg g_3^i \wedge \neg g_4^j)]]\\
\wedge & [t_2 Goal_i[\bar{t}, g_1^i] \wedge Goal_i[\bar{t}, g_3^i]] & \wedge & [t_2, Pref_i([\bar{t}, g_3^i], [\bar{t}, g_1^i])]\\
\wedge & [t_2, Bel_i[t_2, Goal_j[\bar{t}, g_2^j]]] & \wedge & [t_2, Bel_i[t_2, \neg Goal_i[\bar{t}, g_4^j] \vee (Goal_i[\bar{t}, g_4^j] \wedge Pref_j([\bar{t}, g_2^j], [\bar{t}, g_4^j]))]]\\
\end{array}$$
$$\rightarrow [t_2, Int_i([\bar{t}, Do(i,\alpha)], [t_3, \neg Do(j,\beta)]) \wedge Int_i([\bar{t}, Do(i,\alpha)] \rightarrow [t_3, \neg Do(j,\beta)])]$$
$$\wedge [t_2, Int_i([t_3, \neg Do(j,\beta)], g_3^i) \wedge \neg Goal_i[\bar{t}, g_1^i] \wedge Send_{ij}Accept([\bar{t}, Do(i,\alpha)])]$$

Figure 1: Threats production axiom (A1) and threat evaluation axiom (A2). *Credible* stands for an axiom that j will use for estimations whether β is a credible threat for i to do α. *Appropriate* stands for axioms that will specify how to choose β, when several such βs exist.

The teenager refuses on the grounds that she may suffer bad health from staying up late. The parent points out that the teenager had stayed up until 2 a.m. for a party the previous week, without suffering from any ill-effects, and brings it up as a counterexample for the argument.

Following is an example from the robots on Mars. Suppose, R_h requests R_m to make a survey of the terrain using its navigation skills. R_m's temperature sensors indicate that in some areas there may be high temperature pockets and these may bring harm to its electronic circuitry. R_m refuses on these grounds. R_h points out that in the past two days, R_m has withstood much higher temperatures created during the explosions used in the digging process, without any evidence of harm to its circuitry. R_h brings it up as a counterexample to convince R_m to undertake the survey.

Appeal to "Prevailing Practice"

In this case, j gets a refusal from i to do α on the grounds that it contradicts goal g of i. If j believes that another agent k had done the same α and it did not contradict the same goal g held by k at the time, it uses it as an argument. For example, a teacher intends that a student talented in baseball should stay after school for extra lessons. This will contribute to the teacher's desire to build a good baseball team at school. He asks the student to do so but the student refuses on the grounds that this will adversely affect his academic performance. The teacher points out that last year the star baseball player of the class was also an A* student, that several good players have also been good students, and encourages the student to take up the activity.

With the robots on Mars, consider the following mention of prevailing practice in an argument. As in the counterexample scenario, R_h requests R_m to make a survey of the terrain using its navigation skills. R_m's temperature sensors indicate that in some areas there may be high temperature pockets and these may bring harm to its electronic circuitry. R_m refuses on these grounds. R_h points that both itself and R_e was exposed to much higher temperatures two days ago, and had withstood them quite well.

Appeal to Self Interest

In this case j believes that α implies one of i's desires. It uses it as an argument. This is a useful argument when j believes that i is not aware of the implication. For example, an employee has a goal to study Japanese, but wants to save money as well. She intends that her company pay for the Japanese lessons and asks the company. The company refuses. The employee points out that having an employee with knowledge of Japanese is a great asset to the company especially in the coming years when the company will face stiff competition from the Japanese. [17]

On Mars, suppose R_e and R_h both plan to dig at site X on Tuesday. If they both dig at the same site, clearly, there will be destructive interference, leading to a malfunctioning of the procedures of either. R_e makes a proposal to R_h to divide their digging time so that R_e digs in the morning while R_h digs in the evening. R_h refuses, since obviously, this proposal reduces its throughput. However, R_e points out, that if R_h refuses, it will result in near zero throughput for R_h and R_e. Instead, sharing the time will further R_h's self-interest much better, since getting half the work done is better than getting no work done.

6.1 Usage of the Formal Axioms

The argumentation axioms can be used in two ways: One use is as a specification for the designer of automated agents for multi-agent environments. The translation from a data-base containing beliefs, intentions, goals and desires into the model representation is simple. There is a need to specify the truth-set of each formula (i.e., the time lines in which the formula is correct). Here, the axioms can be used to build the system and also to check its behavior.

A more appealing use is in assuming that the agents themselves will use the axioms. For example, if an agent derives "$Do(\alpha, i)$", it would try to perform α. Similarly, the agents would use the axioms to evaluate messages, to send argumentations and to update their knowledge bases.

[17] In another setting, agent j requests agent i to give up its intention to do α by pointing out that either α or its side-effect β contradicts one of i's desires.

6.2 Selecting Arguments by the Agent's Type

What can provide guidelines for the agent on which category of argument to use? Sometimes, the belief about the opponent's type may influence the argumentation. If an agent is memoryless, non-bounded threats or future rewards are not applicable[18]. Suppose agent j asks a memoryless agent k to do α and threatens it with β. Let us assume that β is expensive to j. If k won't do α, there is no benefit for j in carrying out the threat (i.e., do β). The only reason that j may do β is to maintain its credibility. However, if agent k doesn't have a memory of past encounters, it has no notion of credibility. But then again, if it is clear that j won't carry its threat (or will keep its promise for future reward) there is no sense in making threats to start with. It seems that in case of memoryless agents only bounded threats or rewards would make sense.

On the other hand, the counterexample argument is appropriate in the case of a memoryless agent. In this case the agent doesn't remember the counterexample, and the purpose of the argument is to bring it to its notice.

Counterexamples may also be useful as an argument for an agent that is not omniscient. This agent may not have realized in the past that its action contradicted its goal. However, the non-omniscient agent may respond with a counter-argument that had it realized the implication, it wouldn't have taken the action in the past either.

Appeal to self-interest is more appropriate in case where the agent is not omniscient or in cases when the agent's beliefs are incomplete. In both cases the agent may not be aware of its self-interest, and such an argument may change its intentions.

7 Conclusions

In this paper we have presented a formal framework for argumentation. A formal mental model of the agents based on possible worlds (time lines) is built. A formal axiomatization scheme has been constructed for argument generation and evaluation based on argument types identified from human negotiation patterns. Future work agenda includes investigating the relations between different modalities, as well as the change in the modalities over time as the result of the argumentation process and events and observations from the environment. An analysis of the credibility and reputation of adversaries based on repeated encounters is also being incorporated into the argumentation process.

References

[1] H. Abelson. *Persuation.* Springer Pub. Co., New York, 1959.

[2] M. E. Bratman. What is intention? In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 15–31. MIT Press, 1990.

[3] M. E. Bratman. Shared cooperative activity. *The Philosophical Review*, 1992. Forthcoming.

[4] B. F. Chellas. *Modal Logic: An Introduction.* 1980.

[5] P. Cohen and H. Leveque. Intention is choice with commitment. *Artificial Intelligence*, 42:263–310, 1990.

[6] Jon Doyle. Rationa belief revision. In *Proc. of KR-91*, 1991.

[7] B. Grosz and S. Kraus. Collaborative plans for group activities. In *IJCAI93*, French, 1993.

[8] S. Kraus, M. Nirkhe and K. Sycara. Reaching agreements through argumentation: a logical model. In *Proc. of the 12th International workshop on Distributed AI*, May 1993.

[9] G. Kiss and H. Reichgelt. Towards a semantics of desires. In *Proc. of the Third European Workshop on Medeling Autonoumous Agents in a Multi Agent World*, Germany, 1991.

[10] Kurt Konolige. On the relation between default and autoepistemic logic. *Artificial Intelligence*, 35:343-382, 1988.

[11] S. Kraus and J. Wilkenfeld. Negotiations over time in a multi agent environment: Preliminary report. In *Proc. of IJCAI-91*, pages 56–61, Australia, 1991.

[12] J. a. P. Martins and S. C. Shapiro. A model for belief revision. *Artificial Intelligence*, 35(1):25–79, 1988.

[13] M. Nirkhe, S. Kraus, and D. Perlis. Situated reasoning within tight deadlines and realistic space and computation bounds. In *Proceeedings of the Second Symposium On Logical Formalizations Of Commonsense Reasoning*, 1993.

[14] M. E. Pollack. Plans as complex mental attitudes. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 77–103. MIT Press, 1990.

[15] A. S. Rao and M. P Georgeff. Asymmetry thesis and side-effect problems in linear-time and branching-time intention logics. In *Proc. of IJCAI-91*, pages 498-504, Australia, 1991.

[16] Y. Shoham. Agent oriented programing. *Artificial Intelligence*, 60(1):51–92, 1993.

[17] K. Sycara. Persuasive argumentation in negotiation. *Theory and Decisions*, 28:203–242, 1990.

[18] K.P. Sycara. *Resolving Adversarial Conflicts: An Approach to Integrating Case-Based and Analytic Methods.* PhD thesis, School of Information and Computer Science, Georgia Institute of Technology, 1987.

[19] Becky Thomas, Yoav Shoham, Anton Schwartz, and Sarit Kraus. Preliminary thoughts on an agent description language. *International Journal of Intelligent Systems*, 6(5):497–508, August 1991.

[20] M. Vardi. On the complexity of epistemic reasoning. In *Proceedings of the 4th Annual Symposium on Logic in Computer Science*, 1989.

[21] M. Wellman and J. Doyle. Preferential semantics for goals. In *Proc. of AAAI-91*, California, 1991.

[22] Eric Werner. Toward a theory of communication and cooperation for multiagent planning. In *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 129–143, Pacific Grove, California, March 1988.

[18] Note, that this discussion is specific to automated agents. Human negotiators always have at least some memory of the past.