To appear in *Journal of Philosophical Logic, 1993*

A Logic of Intentions and Beliefs[*,†,‡]

Munindar P. Singh   Nicholas M. Asher
Center for Cognitive Science
The University of Texas
Austin, TX 78712
USA

msingh@cs.utexas.edu

Abstract

Intentions are an important concept in Artificial Intelligence and Cognitive Science. We present a formal theory of intentions and beliefs based on Discourse Representation Theory that captures many of their important logical properties. Unlike possible worlds approaches, this theory does not assume that agents are perfect reasoners, and gives a realistic view of their internal architecture; unlike most representational approaches, it has an *objective* semantics, and does not rely on an *ad hoc* labeling of the internal states of agents. We describe a minimal logic for intentions and beliefs that is sound and complete relative to our semantics. We discuss several additional axioms, and the constraints on the models that validate them.

[*] This is a considerably extended and revised version of a paper entitled *"Towards a Formal Theory of Intentions"* that appears in the proceedings of the European Workshop on Logic in AI.

[†] We are indebted to Allen Emerson and Rob Koons and to two anonymous referees for comments.

[‡] This work was partially supported by the Microelectronics and Computer Technology Corporation, and by the National Science Foundation (through grant # IRI-8945845 to the Center for Cognitive Science, University of Texas).

1 Introduction

An understanding of intentions is important to several subfields of Artificial Intelligence (AI), especially, speech act theory [3, 4, 10, 12], discourse processing [18], planning [17], and plan recognition [2, 24, 28]. We present a formal theory of intentions and beliefs that is based on Discourse Representation Theory (DRT) [5, 6, 19]. Our theory involves a formal model of time and possibility, and also explicitly models the structure of the agents' internal states.

Before we turn to the presentation of our theory, we must closely examine what one desires, or ought to desire, from a theory of intentions from the standpoint of AI. This, of course, depends on what one might want to do with such a theory. A theory of intentions is needed at the foundational level of study in AI and Cognitive Science in order to complete an account of intelligent and, possibly, rational agency [13, 31]. Intentions are important attitudes of intelligent agents and, for resource-bounded agents, cannot be reduced to simple considerations of the optimality of decisions. Their importance to cognition and their use in AI has been defended extensively in the literature [4, 8, 11, 10, 12, 17, 18, 30]. The roles of intentions mentioned below are especially important when one is interested in agents who would not otherwise (because of their limitations) be able to make appropriate or rational decisions. Intentions, and therefore a theory of intentions, is needed so that

- Designers and analyzers may abstractly characterize the behavior they expect from the agents they are, respectively, designing and analyzing; and
- Agents so designed may interact intelligently with each other, i.e., cooperate with others, if they are cooperative, or compete successfully against them, if they are not.

A theory of intentions is also necessary in order to make sense of talk in AI about "plans." Plans are mostly treated operationally in AI. A theory of intentions can provide a principled substitute for areas like Natural Language Understanding, where the plans and intentions of agents must be understood, in order to

- Fully understand their utterances
- Communicate effectively with them, i.e., generate felicitous utterances, say, in replying to their queries
- Understand descriptions of their actions (this is called "story understanding")
- Provide assistance to them (this is important in the design of user interfaces, and in Computer Assisted Instruction)

These applications of the theory of intentions impose certain requirements on it. A useful theory would

- Provide an abstract account of the architecture of intelligent agents, especially with regard to their beliefs and intentions; this account would serve as the foundation for the semantic model incorporated in the theory

- Validate some general inferences involving intentions
- Provide for several definitions of intentions, each corresponding to a different species of agent, as might be encountered in different applications
- Provide a connection to events and plans
- Provide a connection to the structure of discourses

We see the work reported here as a step towards the greater goal of a unified theory of cognition, action and communication. Such a unified theory has been evolving in the DRT framework over the last few years. Kamp's thesis of the Unity of Thought and Information [20], and Asher's work on the attitudes [5], and on their relationship with information [7] must be cited in this context. DRT is a useful framework for the general project for several reasons. Firstly, DRT is a theory of discourse meaning that captures many aspects of the information typically encoded in natural language utterances. These aspects are important not only in the AI areas that deal with language directly but also with those that deal with information and action at large. Secondly, the representation structures that DRT posits can very naturally be used to describe attitudes and to connect them to an agent's actions; these structures effectively capture the structure of discourses and events, especially with regard to the treatment of conceptual individuals, and the ways in which they may be anchored to each other and to real individuals in the world. We sketch just enough of DRT in this paper so that the presentation here is self-contained.

A study of the (mostly AI) literature [4, 8, 10, 17, 18, 25, 27, 28, 30] yields the following important properties of intentions. Intentions are about future events. An agent with an intention should believe that it can be realized, at least along some future. If an intention were impossible to achieve, it would be functionally redundant.[1] Thus if an agent *believes* an intention to be futile, he should drop it (or drop the belief). This is a very weak claim, namely, that the agent merely believes that the intended condition is possible. It admits success due to fortuitous circumstances. To use an example due to a referee of this paper, a student can intend to graduate with perfect grades even though she might rate the chances of success as very low. But, on the other hand, if she knows she obtained a C last semester (and is a sufficiently smart agent), she would consider it impossible that she will graduate with perfect grades. Hence, she will not intend to do so.

While we do not consider attitude revision as such, we do wish for our account to be able to take care of different constraints on an agent's attitudes. Agents do not necessarily intend all the consequences of their intentions, or even all the consequences they anticipate. These properties follow naturally from our theory. Furthermore, an agent's intentions "tend" to constrain his further intentions, and tend to persist until they are fulfilled: this allows them to have the functional role (in the agents' lives, as it were) of providing the context

[1] If we exclude deviant cases in which an agent has an impossible intention (say, for p) but, as a consequence of having it, achieves something else (say, q) that is useful to him. In this case, he prefers the scenarios on which q occurs, so the objective part of the model suggests that it is q that he really intends to achieve.

of further reasoning (including the adoption of further intentions), and of simplifying his decision-making. These properties, while not valid in the minimal logic, are expressible in extensions of it (see §5).

2 More Motivations

Our theory, like the so-called sentential theories [21], and unlike most possible-worlds based theories [10], avoids attributing logical omniscience to agents, since it does not require that agents' intentions or beliefs be closed under logical equivalence (thus it also avoids validating closure under logical consequence). At the same time, this theory has advantages over the sentential theories as well. First, it captures the notion of *approximation* that is crucial in the semantics of attitudes like intention and belief. We do not require that a claim of a belief or intention be deemed true only if a corresponding sentence is found in the agent's mind; just that the content of the putative belief or intention approximate the content of (some matching component of) the agent's mental state. This makes it possible for us to assign beliefs and intentions to an agent about another agent's beliefs and intentions, without requiring that the first agent have perfect knowledge of the second agent's cognitive state.

The proposed approach yields a weak logic for intention and belief that we describe in §4. We take this logic as characterizing the minimal rationality that our agents must exhibit for it to make sense for us to ascribe beliefs and intentions to them. Further, as we show in §5, the algebraic structure of the DRS's allows us to establish a variety of closure conditions for intentions and beliefs to capture different logics, and to do so in a semantically and pragmatically felicitous manner. Thus our approach also avoids the charges of *ad hoc*-ism often levied against the sentential approaches, e.g., by Levesque [23], and Fagin & Halpern [15]. The main advantage of our approach is that it allows us to model the internal architecture of intelligent agents far more realistically than the other formal approaches can. As a result, we can exhibit with ease the interconnections that may exist between different attitudes, and also the anchoring of the attitudes to the real world. These interconnections and anchors are crucial in the formulation of plans, and in reasoning about plans and actions.

Now for some general intuitions. A semantics of attitudes assigns truth to a putative attitude just when it correctly characterizes the internal state of the agent. In turn, the question of whether a putative attitude correctly characterizes an agent's internal state must be answered in the framework of a general architecture of agents, and their relation to the world. Our assumption is that DRS's may serve as fair external characterizations of the agents' cognitive states (we do not claim that DRS's are actually *present* as sentences or quasi-sentences in the agents' minds). The internal state of an agent determines, or at least influences, his reasoning and his choice of actions. The actual consequences of his choices depend on what the world is like. The internal state of a well-attuned agent will be connected fairly tightly to his environment. These connections are in terms of the way in which parts of his state are *anchored* to parts of the world—these parts may be individuals or propositions. For example, a frog may at some point be said to believe that the fly it has been pursuing is within tongue range; it may be said to intend to eat that fly; or it may have a belief that it is raining (and therefore croak for a mate). Frogs are successful as a species only because quite often the flies they believe to have within range are real flies that are actually within range (i.e., conceptual flies are anchored to real flies), and when the part of their state that governs croaking is *on*, it is actually raining (i.e., conceptual conditions are anchored to real conditions in the world, perhaps via conceptual individuals).

It should be clear that the meaning of beliefs and intentions derives not just from their interrelationships, but also from how they connect to the agent's environment. Indeed, if it were not for these external anchors, attitude ascription would reduce to the futile game of guessing the internal structure of different agents, and the design of agents to the process of arbitrarily labeling their internal states. Given two agents whose cognitive states can be described by the same DRS's, but whose referents are anchored differently, we would assign different beliefs to them. For example, an agent sweating in Phoenix would have beliefs about the weather in Phoenix, while an agent with the same cognitive state in Austin would have beliefs about the weather in Austin. That anchoring is important in giving the semantics of beliefs is a claim that we take as true. DRT, by itself, is used merely as a technical framework and has nothing to say about whether such anchoring is important. As will become clear in §3, anchors are captured by *embedding functions* in the semantics. If we wish to ignore the anchoring, we can just use an embedding function whose domain is the empty set. This would make, in the above example, the agents in Phoenix and Austin, respectively, come out as having the same beliefs.

It must, however, be remarked that it is not acceptable to ignore the internal structure of the agent entirely; the choices made by an agent, and his (most especially, verbal) behavior do not just depend on the anchoring of his internal state in the world, but on its structure as well: Kripke's 'London' versus 'Londres' example [22], discussed in a DRT framework by Asher [5, pp. 142–143] and by Kamp[20, pp. 253–254], is a case in point. Kripke develops a convincing story in which a character, Pierre, ends up with contradictory beliefs about 'London' and 'Londres,' respectively, even though they are both anchored to the same metropolis in England. Pierre is to be distinguished from a truly confused person who has the same beliefs about the city of London, but with only one internal referent. In considering the structure of the agents' internal states, our approach differs from classical possible worlds approaches; in considering external anchors, it differs from the sentential approaches; and by considering both structures and anchors, it successfully applies to the continuum of intelligence from frogs to humans. We will come back to this point in §3.4.

The semantic conditions for beliefs and intentions are a simplification of the ones given by Asher in his "complete theory" for the case of beliefs [5, pp. 171-173]. This simplification results in part because we consider an explicit assignment function assigning cognitive states to agents. As a result, we have also been able to separate out the components of *content* and *honesty*, yielding a more perspicuous analysis of beliefs and intentions. We have also been able to consider some of the interactions between beliefs and intentions. This is important since it brings us closer to the ultimate goal of a unified theory of actions, beliefs and intentions. The theory presented here is a theory of beliefs and intentions, *not* of belief and intention reports—a theory of belief reports being a contribution of Asher [5]. It considers the logical aspects of these concepts and the consequences of making different assumptions about the model. These aspects and consequences underlie a theory of belief and intention reports, but are distinct from it.

In §3, we present the formal language and model. In §4, we motivate a minimal logic for intentions and beliefs. In §5, we list some important extensions to the basic logic in terms of axioms and and the constraints on models in which they are validated.

3 Formal Language and Formal Semantics

Our sentences (DRS's) [5] are members of the language, **DRS**, generated by the following semi-formal grammar. The temporal part of the grammar is inspired by CTL*, with the addition of the "sometimes in the past" operator, P[14].

1. *DRS* $\longrightarrow$
   (a) $\langle U, \mathit{Cond}\rangle$ |
   (b) *predicate(variable, ..., variable)* |
   (c) $\neg$ *DRS* |
   (d) *DRS* $\vee$ *DRS* |
   (e) *DRS* $\rightarrow$ *DRS* |
   (f) *variable* Believes *DRS* |
   (g) *variable* Intends *DRS* |
   (h) P*DRS* |
   (i) A*DRS'* |
   (j) E*DRS'*
2. *DRS'* $\longrightarrow$
   (a) *DRS* |
   (b) *DRS'* U*DRS'*
3. *U* $\longrightarrow$ *list of variables*
4. *Cond* $\longrightarrow$ *list of DRS*

As usual, $\mathsf{F}p$ abbreviates $\mathsf{trueU}p$, and $\mathsf{G}p$ abbreviates $\neg\mathsf{F}\neg p$. U_K is the "universe" of DRS K, Cond_K its "conditions set," and U_K^* its "extended universe" that contains all the variables in all its sub-DRS's. Often, "condition" is used for "DRS." We stipulate that no variable be redeclared—at worst, this requires a renaming of variables. "$K \wedge L$" abbreviates $\langle U_K \cup U_L, \mathrm{Cond}_K \cup \mathrm{Cond}_L\rangle$. Clearly, "$\wedge$" is idempotent, commutative and associative, as it should be. In the sequel, "$\wedge$" is sometimes applied to sets of DRS's.

$M = \langle \mathbf{W}, \mathbf{T}, <, \mathbf{I}, \mathbf{A}, \mathbf{C}, [\![\,]\!] \rangle$ is a model. Here $\mathbf{W}$ is a set of possible worlds; $\mathbf{T}$ is a set of possible times; $<$ is a partial order; $\mathbf{I}$ is a set of individual objects; $\mathbf{A} \subseteq \mathbf{I}$ is a set of agents; $\mathbf{C}$ is a class of functions assigning cognitive states to the agents at different worlds and times, i.e., $(\mathbf{W} \times \mathbf{T}) \mapsto (\mathbf{A} \mapsto \mathbf{DRS})$; $[\![\,]\!]$ assigns sets of world-time pairs to each n-tuple of individuals for each n-ary predicate.[2]

Each $w \in \mathbf{W}$ has exactly one history, constructed from the times in $\mathbf{T}$. Histories are partially ordered by temporal precedence ($<$), branch into the future, and are eternal along each branch. Times in the history of a world occur only in the history of that world. A *scenario* at a world and time is any maximal eternal branch starting from the given time. Let $\mathbf{S}_{w,t}$ be the class of all scenarios at world w and time t, and let $\mathcal{S}$ be the class of all scenarios, i.e., the union of $\mathbf{S}_{w,t}$ over all w, t. For discrete histories, scenarios correspond to the *fullpaths* of Emerson [14].

An *embedding function*, f, yields at each $w \in \mathbf{W}$ an *embedding*, f_w, that maps variables to individuals in that world. An embedding, g, *extends* an embedding, h, written $g \sqsupseteq h$, if it agrees with h on the domain of h. Restrictions on embeddings can be directly used to model the anchoring of conceptual individuals onto real ones.

The semantics of the extensional fragment of the language is standard; for the attitudes, two functions *Content* and *Honesty* are combined to give a definition of $\models$ ("satisfaction"). Roughly, the *Content* of an attitude is the set of alternatives it selects. These alternatives are implicit in the attitudes and are known only to us, *qua* theoreticians. Beliefs select the scenarios at whose initial world and time, the believed condition is true under the given embedding; intentions select scenarios that represent courses of events leading to their fulfillment, under the given embedding. Intentions are more complex than beliefs since they are future directed. The contents of several attitudes may be combined. The *Honesty* of an attitude depends on whether it matches structurally with the agent's cognitive state; roughly, it is the class of all pairs of the form: (1) a cognitive state with which the given attitude matches, and (2) a connection between individual variables under which this match occurs.

$M \models_{w,t,f} K$ expresses "M satisfies K under f at w, t." $M \models_{S,f} K$ expresses "M satisfies K under f on scenario S." K is *satisfiable* relative to a set of anchors iff for some M, w, t, and f, $M \models_{w,t,f} K$, where f obeys the given anchors. For external anchors, which are of the form "x is anchored to a," we require $f_w(x) = a$. For internal anchors, which are of the form "$x = y$," we require $f_w(x) = f_w(y)$. K is *satisfiable* iff for some model M, world w and time t, $M \models_{w,t,\emptyset} K$, where $\emptyset$ is the embedding function whose domain, at each world, is empty. A DRS, K is *valid* at M and w iff it is satisfiable at all times in M and w. Validity in a model and validity *simpliciter* may be defined analogously.

[2]Strictly speaking, we ought to put the $\mathbf{C}$ and $[\![\,]\!]$ in the interpretation, and make $\mathbf{C}$ assign not DRS's but model-theoretic counterparts of DRS's (namely, DRS's as algebraic structures). But, since it is clear that the language does not allow DRS's to be referred to, there is no problem here.

3.1 Satisfaction conditions

The satisfaction conditions for $\neg$, $\vee$, $\rightarrow$, and predicates as given below are standard in the DRT literature (and are adapted from those in [5, 6]); the ones for the temporal operators too are standard (and are adapted from those in [14]). The ones for $\mathsf{Believes}$ and $\mathsf{Intends}$ are novel to this paper.

- $M \models_{w,t,f} \psi(x_1, \ldots, x_n)$ iff $\langle w, t\rangle \in [\![\psi]\!](\langle f_w(x_1), \ldots, f_w(x_n)\rangle)$
- $M \models_{w,t,f} K \vee L$ iff $(\exists g : g_w \sqsupseteq f_w \wedge (M \models_{w,t,g} K \vee M \models_{w,t,g} L))$
- $M \models_{w,t,f} \neg K$ iff $\neg(\exists g : g_w \sqsupseteq f_w \wedge M \models_{w,t,g} K)$
- $M \models_{w,t,f} K \rightarrow L$ iff $(\forall g : g_w \sqsupseteq f_w \wedge M \models_{w,t,g} K \rightarrow (\exists h : h_w \sqsupseteq g_w \wedge M \models_{w,t,h} L))$
- $M \models_{w,t,f} \mathsf{E}K$ iff $(\exists g : g_w \sqsupseteq f_w \wedge (\exists S : S \in \mathbf{S}_{w,t} \wedge M \models_{S,g} K))$

  E stands for "in *some* scenario"; i.e., K is true in *some* future of t in world w.
- $M \models_{w,t,f} \mathsf{A}K$ iff $(\exists g : g_w \sqsupseteq f_w \wedge (\forall S : S \in \mathbf{S}_{w,t} \rightarrow M \models_{S,g} K))$

  A stands for "in *all* scenarios"; i.e., K is true in *all* futures of t in world w.
- $M \models_{S,f} K\mathsf{U}L$ iff $(\exists g : g_w \sqsupseteq f_w \wedge (\exists t' : t' \in S \wedge M \models_{S',g} L \wedge (\forall t'' : t'' \in S \wedge t \leq t'' \leq t' \rightarrow M \models_{S'',g} K)))$, where S' and S'' are the suffixes of S at times t' and t'', respectively.
- $M \models_{w,t,f} \mathsf{P}K$ iff $(\exists g : g_w \sqsupseteq f_w \wedge (\exists t' : t' \leq t \wedge M \models_{w,t',g} K))$

  Note that P is reflexive; i.e., p entails $\mathsf{P}p$.
- If K is of the form x $\mathsf{Believes}$ L, or is of the form x $\mathsf{Intends}$ L then the following definition applies.

  $M \models_{w,t,f} K$ iff $(\exists \rho, N : \langle \rho, N\rangle \in \mathit{Honesty}(K, \mathbf{C}_{w,t}(f_w(x))) \wedge$ compatible$(f, \rho) \wedge$ $\mathit{Content}_f(K) \supseteq \mathit{Content}_{f\circ\rho^{-1}}(N))$

  This definition is motivated by the intuitive remarks of §2. As argued there, we would like the semantics of intentions and beliefs to incorporate both (1) the structure of the given agent's cognitive state and (2) its connections to the environment. Requirement (1) is captured by the clause requiring the given intention or belief to match with some portion of the agent's cognitive state. This matching is required to take into account the agent's reasoning from the given intention or belief. Requirement (2) is captured by the clause relating the content of the given intention or belief and the content of the corresponding portion of the agent's cognitive state. As defined in §3.2, the content of an intention or belief is the set of possible futures compatible with it—this gives us an external characterization of the agent's actions. In this way, we can differently judge two intentions or beliefs, if (1) they differ in structure, e.g., p and $(p \wedge q) \vee (p \wedge \neg q)$, or (2) they differ objectively in content, e.g., p and $p \vee q$.

In the above definition, $f \circ \rho^{-1}$ must yield a well-defined embedding function—this is what the predicate 'compatible' captures. Compatible(f, ρ) iff $(\forall w, z : w \in \mathbf{W} \wedge z \in \text{Domain}(f_w) \cap \text{Domain}(\rho) \rightarrow (\exists g : g \sqsupseteq f \wedge g_w(\rho(z)) = g_w(z)))$. This ensures that the embedding of DRS N above is not incoherent. This compatibility condition is weaker than the one given by Asher, who requires that ρ be a 1-1 mapping [5, p. 173]. Thus this definition is a generalization of his. *Content*, *Honesty* and $\circ$ are discussed in §3.2, §3.3 and §3.4, respectively.

- $M \models_{S,f} K$ iff $(\exists w, t : S \in \mathbf{S}_{w,t} \wedge M \models_{w,t,f} K)$
- $M \models_{w,t,f} K$ iff $(\exists g : g_w \sqsupseteq f_w \wedge U_K \subseteq \text{Domain}(g_w) \wedge (\forall C \in \text{Cond}_K : M \models_{w,t,g} C))$
- $M \models_{w,t} K$ iff $(\exists f : M \models_{w,t,f} K)$

3.2 Content

The *Content* function must be defined for all DRS's of the form $x\,\mathsf{Believes}\,K$ or $x\,\mathsf{Intends}\,K$. One component of the semantics of beliefs and intentions has to do with the courses of events or scenarios compatible with them. These are the scenarios over which the given belief or intention is satisfied. Intentions and beliefs, especially the former, can be seen as having a component of meaning that has to do with the agent's dispositions to act in certain ways. How agents act depends on numerous factors, but if one sees an intention as specifying an abstract action, that action is done successfully on precisely the scenarios that are compatible with it (i.e., with the corresponding intention). Hence, the importance of the notion of content.

- Roughly, the content of $x\,\mathsf{Believes}\,K$ is the set of scenarios at whose initial world and time, K is true under the given embedding. We include all possible scenarios at each world and time just to facilitate combination with the contents of intentions.

  $Content_f(x\,\mathsf{Believes}\,K) = \{S | (\exists w, t : S \in \mathbf{S}_{w,t} \wedge (\exists g : g_w \sqsupseteq f_w \wedge M \models_{w,t',g} K))\}$

- The content of $x\,\mathsf{Intends}\,K$ is the set of scenarios such that if the world developed along any of them, the intention would succeed; i.e., K would eventually become true. This intuition can be formalized as follows:

  $Content_f(x\,\mathsf{Intends}\,K) = \{S | (\exists w, t : S \in \mathbf{S}_{w,t} \wedge (\exists t', g : t' \in S \wedge g_w \sqsupseteq f_w \wedge M \models_{w,t',g} K))\}$

As a consequence of this definition, the following properties of intentions are accounted for: (1) an agent with an intention tacitly considers it possible that his intention will be fulfilled (this is motivated in [25]), and (2) is tacitly restricted by his intention to scenarios in which it is achieved (this is motivated in [8]). These are two of the most important properties of intentions.

The content of any set of attitudes is the intersection of their respective contents. Let $Content_f^I$ denote the function that picks out, and gives the content of, the subset of intentions of its argument; let $Content_f^B$ be the corresponding function for beliefs. The content (relative to an embedding function) of a DRS whose conditions set contains only DRS's of the form $x\ \mathsf{Believes}\ K$ or $x\ \mathsf{Intends}\ K$ is the content of its conditions set, relative to a function that extends the original function, and whose Domain includes the universe of the given DRS.

3.3 Honesty

The relation of the structure of a putative attitude to that of the agent's actual internal state is important. The honesty of a putative attitude K is given by the set of pairs of the following form: one component is a renaming of the variables in K; the other component a DRS, L, that "subsumes" K under that renaming (the relations of subsumption, $\preceq^I$ and $\preceq^B$, are described in detail in §3.4). There is a component of the meaning of an intention or belief that has to do with the choices an agent with such an intention or belief would make. In general, a number of factors influence this choice, but one contributor is the structure of the agent's cognitive state. Even though a given intention or belief may be externally identical to an intention or belief the agent has, the agent's cognitive state may not be properly characterized by it. For example, a true belief that a certain 100-digit number is prime differs from a belief that $2 + 2 = 4$, even though they are both true in all possible worlds.

It is convenient to relativize Honesty to a cognitive state (formally, a DRS—called "CS" here) from whose condition set the DRS's L are selected. The set of pairs alluded to above thus enumerates the possible "connections" that must be made between the attitude and the given cognitive state. As a consequence of this definition, the following properties automatically hold in intention and belief contexts: (1) left and right simplification of $\wedge$, (2) commutativity of $\wedge$, and (3) existential generalization (see §4 below). Formally, we have the following definitions. Here we consider all π's, even those whose domain properly includes $\{x\} \cup U_K^*$.

- $Honesty(x\ \mathsf{Believes}\ K, \mathrm{CS}) = \{\langle \pi, L\rangle | L \in \mathrm{Cond}_{\mathrm{CS}} \wedge (x\ \mathsf{Believes}\ K) \preceq_\pi^B L\}$
- $Honesty(x\ \mathsf{Intends}\ K, \mathrm{CS}) = \{\langle \pi, L\rangle | L \in \mathrm{Cond}_{\mathrm{CS}} \wedge (x\ \mathsf{Intends}\ K) \preceq_\pi^I L\}$

3.4 Subsumption Conditions

As remarked in §1, we intend our theory to apply to a range of intelligent agents from frogs to humans. The former do not seem to do any symbolic reasoning, so we must be careful in applying a DRS-based theory on them. The basic difference between frogs and humans, when treated as intelligent agents, is that the former have biologically determined, and rather rigid forms of behavior. Unlike a human, a frog (let us stipulate) represents and distinguishes between a small number of conditions (e.g., hungry versus satisfied, raining versus dry), and represents a small number of conceptual individuals, e.g., a fly to eat, a predator to avoid, a potential mate to attract.[3] Thus a frog is able to effectively have only a limited set of beliefs and intentions. The honesty of a putative attitude of a frog depends on how it is related to this limited set. This would suggest that honesty is best characterized in a species-relative manner: that is indeed the case.

Formally, we now define two relations, $\preceq^I$ and $\preceq^B$, which reflect some restrictions on the structures of our agents' internal states, and the reasoning power that they are endowed with; e.g., $K_1 \preceq^B K_1 \wedge K_2$ means that an agent can perform left simplification on his beliefs. The relation $\preceq^I$ is meant to apply to intentions and the relation $\preceq^B$ to beliefs—the differences between these relations reflect the different functional roles that intentions and beliefs play in an agent's life. Note that $\preceq^I$ and $\preceq^B$ (and honesty, in general) do not apply to the semantics of all conditions, but only to the semantics of attitudes. For frogs, these relations would be almost empty; for perfect reasoners, they would allow all valid deductions. Let π, ρ and σ be *alphabetic functions*, which rename variables. These functions allow us to model the connections among DRS's. An alphabetic function ρ extends another alphabetic function π, or $\rho \sqsupseteq \pi$, iff for all $a \in \text{Domain}(\pi)$, $\rho(a) = \pi(a)$. Define $\rho = \rho_1 \cup \rho_2$ as the alphabetic function such that $\text{Domain}(\rho) = \text{Domain}(\rho_1) \cup \text{Domain}(\rho_2)$ and $(\forall z : z \in \text{Domain}(\rho_1) \rightarrow \rho(z) = \rho_1(z))$ and $(\forall z : z \in \text{Domain}(\rho_2) \rightarrow \rho(z) = \rho_2(z))$. Clearly, this is well-defined only if ρ_1 and ρ_2 do not differ on any z. We now define $\preceq^I$ and $\preceq^B$ relative to an alphabetic function (we write $\preceq^*$ in conditions that apply to both).[4]

1. $K \preceq^*_\pi L$ if $(\forall K' : K' \in \text{Cond}_K \rightarrow K' \preceq^*_\pi L)$
2. $K \preceq^*_\pi L$ if $(\exists L' : L' \in \text{Cond}_L \wedge K \preceq^*_\pi L')$
3. $\psi(x_1, \ldots, x_n) \preceq^*_\pi \psi(\pi(x_1), \ldots, \pi(x_n))$
4. $\mathsf{T}K \preceq^*_\pi \mathsf{T}L$ if $K \preceq^*_\pi L$

   Here T may be any one of E, A or P.
5. $(K_1 \mathsf{U} K_2) \preceq^*_\pi (L_1 \mathsf{U} L_2)$ if $(K_1 \preceq^*_\pi L_1$ and $K_2 \preceq^*_\pi L_2)$
6. $\neg K \preceq^*_\pi \neg L$ if $L \preceq^*_\pi K$
7. $(K_1 \rightarrow K_2) \preceq^*_\pi (L_1 \rightarrow L_2)$ if $(L_1 \preceq^*_\pi K_1$ and $K_2 \preceq^*_\pi L_2)$ and $(\forall x : x \in U_{K_1} \rightarrow \pi(x) \in U_{L_1})$

   The latter condition is needed because the semantics of $\rightarrow$ (as given in §3.1) gives the variables in the antecedent a universal quantifier reading.
8. $(K_1 \vee K_2) \preceq^*_\pi (L_1 \vee L_2)$ if $(K_1 \preceq^*_\pi L_1$ and $K_2 \preceq^*_\pi L_2)$ or $(K_1 \preceq^*_\pi L_2$ and $K_2 \preceq^*_\pi L_1)$

   Commutativity of $\vee$ under both intention and belief.

---

[3] We can thus use conceptual individuals to model the "indexical-functional" aspects of the environment [1]. This approach allows us to also make sense of higher level notions like belief and intention.

[4] These definitions are only two of several possible ones, which model agents of different levels of "smartness"—we include these only to make our proposal complete in terms of specifying one kind of limited rational agents, not to present a prescriptive view of rationality. Further variations are considered in §5.

9. $(x\ \mathsf{Intends}\ K) \preceq^*_\pi (\pi(x)\ \mathsf{Intends}\ L)$ if $K \preceq^I_\pi L$

   Since K and L are objects of intentions, their relationship must be appropriate for intentions.

10. $(x\ \mathsf{Believes}\ K) \preceq^*_\pi (\pi(x)\ \mathsf{Believes}\ L)$ if $K \preceq^B_\pi L$

    Since K and L are objects of beliefs, their relationship must be appropriate for beliefs.

11. $(x\ \mathsf{Believes}\ K) \preceq^B_\pi (\pi(x)\ \mathsf{Intends}\ L)$ if $K \preceq^I_\pi \mathsf{EF}L$

    This captures the requirement that all agents explicitly believe their intentions to be fulfillable along some future course of events. See axiom (WA) in §4 below.

We write $K \preceq^* L$, if $(\exists\pi : K \preceq^*_\pi L)$. The two subsumption relations thus characterize two simple logics that an agent (at least, tacitly) uses under intention and belief contexts, respectively.

3.5 Some Lemmas

Some useful properties that are used in the proofs in §5 are given by the following lemmas.

Lemma 1 Let A, B, C and D be pairwise disjoint sets of referents. Let $\kappa : A \mapsto B$, $\mu : C \mapsto D$ and $\rho = \kappa \cup \mu$ be alphabetic functions. Now for any embedding function f, compatible(f, ρ) iff compatible(f, κ) and compatible(f, μ).

Proof

By the definition of compatible in item 3.1 of §3.1, we have the following: $(\exists g' : g' \sqsupseteq f \wedge (\forall w, x : x \in \mathrm{Domain}(f_w) \cap A \rightarrow g'(\kappa(x)) = g'(x)))$ and $(\exists g'' : g'' \sqsupseteq f \wedge (\forall w, x : x \in \mathrm{Domain}(f_w) \cap C \rightarrow g''(\mu(x)) = g''(x)))$. Since A and C are disjoint, ρ is well-defined. Now define an embedding function $g \sqsupseteq f$ such that $(\forall x : x \in \mathrm{Domain}(f_w) \cap A \rightarrow g(x) = g'(x) = g'(\kappa(x)))$ and $(\forall x : x \in \mathrm{Domain}(f_w) \cap C \rightarrow g(x) = g''(x) = g''(\mu(x)))$. Since B and D are disjoint, g is well-defined (i.e., for no x and y is $\kappa(x) = \mu(y)$). But $(\forall x : x \in A \rightarrow \kappa(x) = \rho(x))$ and $(\forall x : x \in C \rightarrow \mu(x) = \rho(x))$. Thus we have that $(\forall w, x : x \in \mathrm{Domain}(f_w) \cap (A \cup C) \rightarrow g(\rho(x)) = g(x))$ or compatible(f, ρ). The converse direction is even simpler: let g' and g'' be g. Thus for any embedding function f, compatible(f, ρ) iff compatible(f, κ) and compatible(f, μ).

Lemma 2 Let N be a DRS, R a set of reference markers, $f : R \mapsto \mathbf{I}$ an embedding and $\rho : R \mapsto U_N$ an alphabetic function. Then $U_N \cap R = \emptyset$ and compatible(f, ρ) implies that $\mathit{Content}_f(N) \supseteq \mathit{Content}_{f\circ\rho^{-1}}(N)$.

Proof

Let w, t, g be such that $g \sqsupseteq (f \circ \rho^{-1})$ and $M \models_{w,t,g} N$. Compatible(f, ρ) ensures that $(f \circ \rho^{-1})$ is a well-defined function whose Domain is a subset of U_N, and is therefore disjoint with R. Thus $(\exists h : h \sqsupseteq g \wedge h \sqsupseteq f \wedge (\forall z : z \in U_N \rightarrow h(z) = g(z)))$. Thus $M \models_{w,t,h} N$. Thus, by the definition of $\mathit{Content}$, $S \in \mathit{Content}_{f\circ\rho^{-1}}(N) \Rightarrow S \in \mathit{Content}_f(N)$, which proves the lemma.

4 A Logic for Intention and Belief

We have so far presented a model theoretic approach to giving the semantics of intentions and beliefs. However, for many purposes in several important subfields of AI, it would be useful to also have a logic that corresponds to the above semantics. These purposes include (1) reasoning by an agent to determine his own plans, and to reason about their consequences [17, 25], (2) and to do the same for the plans of others [28], and (3) a principled approach to the design of multiagent intelligent systems [29]. Points (1) and (2) directly relate to speech act planning and discourse understanding [4, 10, 18, 27] as well. Such a logic would also be useful in multiagent systems where agents must reason about each other's intentions and beliefs to effectively negotiate among themselves. We now turn to a deductive system for our semantics.

At his point we can either develop the proof theory in a standard First Order Temporal Logic framework, or in the DRT framework. We prefer the former for simplicity and ease of exposition. The language we use is the usual first order temporal logic language augmented by predicates for belief and intention. While some symbols, e.g., $\rightarrow$, are reused below, their meanings should be clear from the context. The axiomatization is then quite straightforward.

1. (WA): Weak Anticipation.

   $(x \mathsf{Intends}\, p) \rightarrow (x \mathsf{Believes}\, \mathsf{EF} p)$

   Since p occurs on all the scenarios in the content of an intention for p, those scenarios are automatically in the content of the consequent belief. We have explicitly endowed our agents with the reasoning power to make this connection, and have forced their internal states to be structured appropriately.

2. Simplification in Belief Contexts.

   From $x \mathsf{Believes}\, (p \wedge q)$ conclude $x \mathsf{Believes}\, p$

3. Commutativity in Belief Contexts.

   From $x \mathsf{Believes}\, (p \wedge q)$ conclude $x \mathsf{Believes}\, (q \wedge p)$

4. Associativity in Belief Contexts.

   From $x \mathsf{Believes}\, (p \wedge (q \wedge r))$ conclude $x \mathsf{Believes}\, ((p \wedge q) \wedge r)$

5. Simplification in Intention Contexts.

   From $x \mathsf{Intends}\, (p \wedge q)$ conclude $x \mathsf{Intends}\, p$

6. Commutativity in Intention Contexts.

   From $x \mathsf{Intends}\, (p \wedge q)$ conclude $x \mathsf{Intends}\, (q \wedge p)$

7. Associativity in Intention Contexts.

   From $x \mathsf{Intends}\, (p \wedge (q \wedge r))$ conclude $x \mathsf{Intends}\, ((p \wedge q) \wedge r)$

8. Existential Generalization in Belief Contexts.

   From x $\mathsf{Believes}\,\phi(b)$ conclude x $\mathsf{Believes}\,(\exists y : \phi(y))$

9. Existential Generalization in Intention Contexts.

   From x $\mathsf{Intends}\,\phi(b)$ conclude x $\mathsf{Intends}\,(\exists y : \phi(y))$

10. Underlying logic.

    All substitution instances of the theorems of the underlying (temporal) logic, along with modus ponens (from K and $K \rightarrow L$ conclude L) are available.

The last requirement is important since it relativizes our axiomatization to that of the underlying logic. This helps us factor out the well-known parts of the theory and focus on the novel parts of this paper.

The above axiomatization is quite simple. The associativity and commutativity inferences in belief and intention contexts arise since DRS's may consist of *sets* of sub-DRS's. The simplification inferences arise due to the embedding conditions for DRS's, and the way in which *Content* is defined. Schema (WA) above is validated by the definition of *Content* and the special clause in the definition of $\preceq$. As a result, it is clear that the axiomatization is **sound**. **Completeness** too is simple. The proof sketched below adapts the canonical model technique discussed by Chellas [9, pp. 60, 173] for our ends. Let the above logic be called Σ. Define a canonical model $M = \langle \mathbf{W}, \mathbf{T}, <, \mathbf{I}, \mathbf{A}, \mathbf{C}, [\![\,]\!] \rangle$ for Σ as follows:

1. Let the members of $\mathbf{T}$ all be maximally consistent sets of DRS's (the universe of each DRS belongs to $\mathbf{I}$). That is, each such set is itself the Cond of a DRS. Thus the $\wedge$'s are mapped into sets of sub-DRS's—this is suggested by the fact that in the DRS language, $K \wedge L$ abbreviates $\langle U_K \cup U_L, \mathrm{Cond}_K \cup \mathrm{Cond}_L \rangle$ (see §3). Note that several members of $\mathbf{T}$ can be the same set (since the same situation may occur at several points in the model).

2. $\mathbf{T}$ is partially ordered by $<$, which may branch only in the future. Connected components of $<$ must be formed of DRS's that all have the same universe. These components are the worlds, and belong to $\mathbf{W}$. Scenarios can be induced by $<$ straightforwardly. The constraints that ensure the proper functioning of $<$ with respect to the temporal operators are routine. We refer to each DRS in $\mathbf{T}$ as a world-time pair. $[\![\,]\!]$ yields for each n-ary predicate and n-tuple of referents (selected from the universe of a maximally consistent DRS) the world-time pairs whose Cond's include that predicate applied to that n-tuple. Embeddings simply pair off the variables (from a potentially unlimited sequence) with the referents in the universe of each world-time pair. We use $\langle w, t, f \rangle$ to refer to a maximally consistent set of DRS's and an embedding function. Define $Content_f$ as appropriate sets of scenarios, as induced by $<$.

3. The cognitive state assignment, $\mathbf{C}$ is defined for each world-time pair and each agent. The following constraint must be met. If K is of the form x $\mathsf{Believes}\,L$ or x $\mathsf{Intends}\,L$

then $K \in \langle w, t, f\rangle$ iff $(\exists \rho, N : \langle \rho, N\rangle \in \mathit{Honesty}(K, \mathbf{C}_{w,t}(f_w(x))) \wedge$ compatible(f, ρ) $\wedge \mathit{Content}_f(K) \supseteq \mathit{Content}_{f \circ \rho^{-1}}(N))$. This is unambiguous since the axioms 1 through 9 are determined by the definition of $\preceq$ and *Content* and the properties of sets. Such models exists since we can trivially let $\mathbf{C}$ be such that $\mathbf{C}_{w,t}(x) = \langle \{x\}, \{K | K$ is of the form x $\mathsf{Believes}$ L or x $\mathsf{Intends}$ $L\}\rangle$.

Completeness and soundness of Σ follow from the construction of this model; i.e., $M \models_{w,t,f} K$ iff $K \in \langle w, t, f\rangle$. Thus we have the following theorem:

Theorem 3 Σ is sound and complete relative to M.

5 Axioms and Constraints

The logic given above corresponds to the core or minimal definition of intentions. While this prevents all the inferences involving intentions that are invalid in general, it validates too few inferences (this is because few inferences are valid in general). However, for specific applications, and in talking about agents who are more intelligent than the basic agents considered so far, it is important to be able to state further axioms, as well as the assumptions under which they are valid. Barbara Partee notes that the lack of valid axioms involving beliefs (and, by extension, intentions) provides only negative evidence against specific proposals for their semantics [26, p. 95]. We feel that positive evidence may be generated when agents of different architectures and computational power are considered. That is, while no axiom seems to hold in general, it is important methodologically to consider axioms that hold under different conditions. We now turn to these additional axioms, which may also be seen as defining alternative senses of intentions.

These alternative definitions differ from the core definition only in the additional restrictions on the models, i.e., on the contents of intentions and on the cognitive states of agents, that they require. To clarify the key intuitions involved: we are trying to characterize the intentions of several different species of agents. The intentions of each species mean something slightly different, even though they are variations on the same theme. In our formal semantics, these variations emerge as different constraints on the contents and the structures that are associated with cognitive states. For example, in the weakest sense, the intentions of the members of a species could be allowed to be mutually inconsistent; in stronger senses, they may be required to be mutually consistent. While in the weaker senses of intentions, they may be had by agents mutually independently, in stronger senses, the agents may be expected to combine their different intentions. In one case, intentions could be such that the agent who has them is aware of them (i.e., he believes that he has those intentions); in another case, the agent could also be aware of the intentions he does not have. Loosely put, these variations are analogous to the different axioms of knowledge that one may or may not adopt when describing a particular epistemic logic. But since our present approach involves both content and structure, and neither is by itself considered sufficient to characterize the meaning of the intentions of a species, these axioms correspond to constraints on both contents, and structures.[5]

Some important axioms along with constraints corresponding to them are given below. These constraints are mostly formulated to be natural with respect to out informal model. They are not always the weakest possible. As in §3.2, $Content^I$ (respectively, $Content^B$) yields the content of the subset of intentions (respectively, beliefs) of its argument.

1. **Conjunction:**

   The agent is able to put his intentions together. If x intends p and also intends q, then his cognitive state is structured so that he also intends the complex condition of achieving p and q in some arbitrary temporal order.

   $(x \,\mathsf{Intends}\, p) \wedge (x \,\mathsf{Intends}\, q) \rightarrow x \,\mathsf{Intends}\, ((p \wedge \mathsf{P}q) \vee (\mathsf{P}p \wedge q))$

   Theorem 4 The addition of this axiom to Σ makes it sound and complete for models that satisfy the following constraints[6]

   - $[x \,\mathsf{Intends}\, ((p \wedge \mathsf{P}q) \vee (\mathsf{P}p \wedge q))] \preceq_\pi^* K$ if $(x \,\mathsf{Intends}\, p), (x \,\mathsf{Intends}\, q) \preceq_\pi^* K$
   - $(\bigwedge_{K \in \mathrm{Cond}_{\mathbf{C}_{w,t(x)}}} K) \in \mathrm{Cond}_{\mathbf{C}_{w,t(x)}}$ (assuming renaming of variables as needed)

   Proof

   The constraint on contents that is required to validate this axiom is already met by the definition of *Content*. Let $x \,\mathsf{Intends}\, p$, $x \,\mathsf{Intends}\, q \in \langle w, t, f\rangle$. For brevity, let C refer to $\mathbf{C}_{w,t}(f_w(x))$. Then $(\exists \kappa, K : \langle \kappa, K\rangle \in \mathit{Honesty}(x \;\mathsf{Intends}\, p, C) \wedge$ compatible$(f, \kappa) \wedge$ $\mathit{Content}_f(x \,\mathsf{Intends}\, p) \supseteq \mathit{Content}_{f \circ \kappa^{-1}}(K))$ and $(\exists \mu, M : \langle \mu, M\rangle \in$ $\mathit{Honesty}(x \,\mathsf{Intends}\, q, C) \wedge$ compatible(f, μ) $\wedge \mathit{Content}_f(x \,\mathsf{Intends}\, q) \supseteq$ $\mathit{Content}_{f \circ \mu^{-1}}(M))$. Let $N = (\bigwedge_{K \in \mathrm{Cond}_C} K)$. By the definition of $\preceq$ (and since variables are declared only once, as already stipulated in §3), the honesty conditions are equivalent to $x \,\mathsf{Intends}\; p \preceq_{\kappa'} N$ and $x \,\mathsf{Intends}\; q \preceq_{\mu'} N$, where κ' and μ' are obvious modifications of κ and μ to take care of the renaming of variables. Clearly, compatible(f, κ') and compatible(f, μ') iff compatible(f, κ) and compatible(f, μ), respectively. Let $\rho = \kappa' \cup \mu'$. This is well-defined since (1) no variables are redeclared

[5] A possibility not pursued here is to have all the constraints on contents apply in the core definition itself, and to just vary the constraints on the structure. There are three reasons for not doing so. One, the different constraints that may potentially be considered need not be mutually consistent. Two, such an approach would be tantamount to taking a normative stance and insisting that there was one "true" sense of intentions. Three, it is important to show how the constraints on the content and structure are related: the ones on the content provide the semantic justification for the ones on the structure, and the ones on the structure provide pragmatic (in the sense of how an agent would deliberate) basis for the ones on the content.

[6] The second constraint says that an agent can put parts of his cognitive state together.

in DRS's; and (2) we assume that the agent uses the same referent for x, i.e., himself, in both the DRS's, K and M—this is required anyway for the above axiom to apply coherently. Therefore, by the definition of $\cup$ for alphabetic functions in §3.4, the above conditions hold iff $x\,\mathsf{Intends}\,p \preceq_\rho N$ and $x\,\mathsf{Intends}\,q \preceq_\rho N$. We have $[x\,\mathsf{Intends}\,((p \wedge \mathsf{P}q) \vee (\mathsf{P}p \wedge q))] \preceq_\rho N$ iff the first constraint holds. We have $N \in \mathrm{Cond}_C$ iff the second constraint holds. Thus, $\langle \rho, N\rangle \in \mathit{Honesty}([x\,\mathsf{Intends}\,((p \wedge \mathsf{P}q) \vee (\mathsf{P}p \wedge q))], C)$.

But by Lemma 1 of §3.4, compatible(f, κ') and compatible(f, μ'), iff compatible(f, ρ). At the same time, $\mathit{Content}_f(x\,\mathsf{Intends}\,((p \wedge \mathsf{P}q) \vee (\mathsf{P}p \wedge q))) = \{S|(\exists w, t : S \in \mathbf{S}_{w,t} \wedge (\exists t', g : t' \in S \wedge g_w \sqsupseteq f_w \wedge M \models_{w,t',g} ((p \wedge \mathsf{P}q) \vee (\mathsf{P}p \wedge q))))\}$. But this equals the set $\{S|(\exists w, t : S \in \mathbf{S}_{w,t} \wedge (\exists t', t'', g : t', t'' \in S \wedge g_w \sqsupseteq f_w \wedge M \models_{w,t',g} p \wedge M \models_{w,t'',g} q))\}$. And that reduces to $\{S|(\exists w, t : S \in \mathbf{S}_{w,t} \wedge (\exists t', g : t' \in S \wedge g_w \sqsupseteq f_w \wedge M \models_{w,t',g} p))\} \cap \{S|(\exists w, t : S \in \mathbf{S}_{w,t} \wedge (\exists t'', g : t'' \in S \wedge g_w \sqsupseteq f_w \wedge M \models_{w,t'',g} q))\}$, which is simply $\mathit{Content}_f(x\,\mathsf{Intends}\,p) \cap \mathit{Content}_f(x\,\mathsf{Intends}\,q)$. But by the above that is a superset of $\mathit{Content}_{f\circ\kappa^{-1}}(K) \cap \mathit{Content}_{f\circ\mu^{-1}}(M)$, which is a superset of $\mathit{Content}_{f\circ\rho^{-1}}(N)$ (since renaming does not affect the content when the embedding is modified appropriately, as it is here). Hence we have the result.

2. **Consequential Closure Under Beliefs:**

   $x\,\mathsf{Intends}\,p \wedge x\,\mathsf{Believes}\,\mathsf{AG}(p \rightarrow q) \rightarrow x\,\mathsf{Intends}\,q$

   Theorem 5 The addition of this axiom to Σ makes it sound and complete for models that satisfy the following constraints

   - $[x\,\mathsf{Intends}\,q] \preceq_\pi K$ if $[x\,\mathsf{Believes}\,\mathsf{AG}(p \rightarrow q)], [x\,\mathsf{Intends}\,p] \preceq_\pi K$
   - $(\bigwedge_{K \in \mathrm{Cond}_{\mathbf{C}_{w,t(x)}}} K) \in \mathrm{Cond}_{\mathbf{C}_{w,t(x)}}$ (assuming renaming of variables as needed)

   Proof

   Let $x\,\mathsf{Intends}\,p$, $x\,\mathsf{Believes}\,\mathsf{AG}(p \rightarrow q) \in \langle w, t, f\rangle$. For brevity, let C refer to $\mathbf{C}_{w,t}(f_w(x))$. Then $(\exists \kappa, K : \langle \kappa, K\rangle \in \mathit{Honesty}(x\ \mathsf{Intends}\,p, C) \wedge$ compatible(f, κ) $\wedge \mathit{Content}_f(x\,\mathsf{Intends}\,p) \supseteq \mathit{Content}_{f\circ\kappa^{-1}}(K))$ and $(\exists \mu, M : \langle \mu, M\rangle \in \mathit{Honesty}(x\,\mathsf{Believes}\,\mathsf{AG}(p \rightarrow q), C)$ $\wedge$ compatible(f, μ) $\wedge$ $\mathit{Content}_f(x\,\mathsf{Believes}\,\mathsf{AG}(p \rightarrow q)) \supseteq \mathit{Content}_{f\circ\mu^{-1}}(M))$. Let $N = (\bigwedge_{K \in \mathrm{Cond}_C} K)$. Using arguments such as those in the proof of Theorem 4, define κ' and μ' from κ and μ, respectively. Let $\rho = \kappa' \cup \mu'$. Therefore, the above conditions hold iff $x\,\mathsf{Intends}\,p \preceq_\rho N$ and $x\,\mathsf{Believes}\,\mathsf{AG}(p \rightarrow q) \preceq_\rho N$. We have $x\,\mathsf{Intends}\,q \preceq_\rho N$ iff the first constraint holds (otherwise, we can easily construct examples where this condition fails). We have $N \in \mathrm{Cond}_C$ iff the second constraint holds. Thus, $\langle \rho, N\rangle \in \mathit{Honesty}(x\,\mathsf{Intends}\,q, C)$. By simple algebraic manipulations on the definition of *Content* we obtain that $\mathit{Content}_{f\circ\rho^{-1}}(N) \subseteq \mathit{Content}_f(x\,\mathsf{Intends}\,q)$. Thus $x\,\mathsf{Intends}\,q \in \langle w, t, f\rangle$.

3. **Self Knowledge:**

The agent's beliefs about his intentions are true; i.e., if an agent believes he has an intention, he really does.

$x \,\mathsf{Believes}\, (x \,\mathsf{Intends}\, p) \rightarrow x \,\mathsf{Intends}\, p$

Theorem 6 The addition of this axiom to Σ makes it sound and complete for models that satisfy the following constraints

- $Content_f(x \,\mathsf{Believes}\, x \,\mathsf{Intends}\, q) \subseteq Content_f(x \,\mathsf{Intends}\, q)$
- $[x \,\mathsf{Intends}\, p] \preceq^*_\pi K$ if $[x \,\mathsf{Believes}\, (x \,\mathsf{Intends}\, p)] \preceq^*_\pi K$

Proof

Construct a canonical model as before. For brevity, let C refer to $\mathbf{C}_{w,t}(f_w(x))$. For any point $\langle w, t, f\rangle$ in the model, $x \,\mathsf{Believes}\, x \,\mathsf{Intends}\, p \in \langle w, t, f\rangle$ iff $M \models_{w,t,f} (x \,\mathsf{Believes}\, x$ $\mathsf{Intends}\, p)$, which is the case iff $(\exists \rho, N : \langle \rho, N\rangle \in Honesty(x \;\mathsf{Believes}\; x \;\mathsf{Intends}\, p,$ $C) \wedge$ compatible$(f, \rho) \wedge Content_f(x \,\mathsf{Believes}\, x \,\mathsf{Intends}\, p) \supseteq Content_{f\circ\rho^{-1}}(N))$. But this implies (iff the constraint on $\preceq$ given above holds, and using the definition of *Honesty* in §3.3) that $\langle \rho, N\rangle \in Honesty(x \;\mathsf{Intends}\; p,\, C)$ (i.e., with the same ρ and N). And we have $Content_f(x \,\mathsf{Intends}\, p) \supseteq Content_{f\circ\rho^{-1}}(N)$ iff the constraint on contents given above holds. Combining these two results, we have that $(\exists \rho, N : \langle \rho, N\rangle \in Honesty(x$ $\mathsf{Intends}\; p, C) \wedge$ compatible$(f, \rho) \wedge Content_f(x \,\mathsf{Intends}\, p) \supseteq Content_{f\circ\rho^{-1}}(N))$. Thus this axiom is determined by the given constraints.

4. **Positive Introspection:**

   The agent knows what intentions he has. This is the inverse of the self knowledge axiom listed above.

   $x \,\mathsf{Intends}\, p \rightarrow x \,\mathsf{Believes}\, (x \,\mathsf{Intends}\, p)$

   Theorem 7 The addition of this axiom to Σ makes it sound and complete for models that satisfy the following constraints

   - $Content_f(x \,\mathsf{Intends}\, q) \subseteq Content_f(x \,\mathsf{Believes}\, x \,\mathsf{Intends}\, q)$
   - $[x \,\mathsf{Believes}\, (x \,\mathsf{Intends}\, p)] \preceq^*_\pi K$ if $[x \,\mathsf{Intends}\, p] \preceq^*_\pi K$

   Proof

   The proof for this case is a simple variation of that for self knowledge given above.

5. **Deliberate Intentions:**

   The agent really intends to have the intentions he has; i.e., the agent chooses his intentions deliberately.

   $x \,\mathsf{Intends}\, p \rightarrow x \,\mathsf{Intends}\, (\mathsf{A}(x \,\mathsf{Intends}\, p)\mathsf{U} p)$

Theorem 8 The addition of this axiom to Σ makes it sound and complete for models that satisfy the following constraints

- $Content_f(x\,\mathsf{Intends}\,q) \subseteq Content_f(x\,\mathsf{Intends}\,(\mathsf{A}(x\,\mathsf{Intends}\,p)\mathsf{U}q))$
- $[x\,\mathsf{Intends}\,\mathsf{A}(x\,\mathsf{Intends}\,p)\mathsf{U}p] \preceq^*_\pi K$ if $[x\,\mathsf{Intends}\,p] \preceq^*_\pi K$

Proof

The proof mimics the one given for self knowledge above.

6. **Self Control:**

   If an agent intends to achieve a state where he has a particular intention, he can directly take that intention on *now*; i.e., the agent can control his cognitive state.

   $x\,\mathsf{Intends}\,(x\,\mathsf{Intends}\,p)\rightarrow\ x\,\mathsf{Intends}\,p$

Theorem 9 The addition of this axiom to Σ makes it sound and complete for models that satisfy the following constraints

- $Content_f(x\,\mathsf{Intends}\,x\,\mathsf{Intends}\,q) \subseteq Content_f(x\,\mathsf{Intends}\,q)$
- $[x\,\mathsf{Intends}\,p] \preceq^*_\pi K$ if $[x\,\mathsf{Intends}\,(x\,\mathsf{Intends}\,p)] \preceq^*_\pi K$

Proof

Again the proof mimics the one given for self knowledge above.

The above axioms and constraints are not all that can be stated about intentions. An important subclass of axioms includes negative introspection and deliberate non-intentions. These are considered below. These axioms cannot be treated on par with the other axioms since their antecedents are negations of attitude claims (and their consequents are positive attitude claims). In the framework of this paper, an attitude may fail to hold for any of two reasons: (1) it is not honest relative to the agent's real cognitive state; and (2) its content is not a superset of the content of the agent's real cognitive state. Thus the constraints that correspond to these axioms are of the form "[content-condition $\vee$ honesty-condition] $\Rightarrow$ [content-condition $\wedge$ honesty-condition]." Thus these constraints are no longer modular between content and honesty conditions in that the antecedent of each constraint must involve both kinds of conditions. Whenever a constraint requires that a DRS be inserted into the cognitive state, we assume that all the markers declared in it are appropriately renamed.

7. **Negative Introspection:**

   The agent knows what intentions he does not have.

   $\neg x\,\mathsf{Intends}\,p\rightarrow\ x\,\mathsf{Believes}\,\neg(x\,\mathsf{Intends}\,p)$

Theorem 10 The addition of this axiom to Σ makes it sound and complete for models that satisfy the following constraints

- $[(x\,\mathsf{Intends}\,q \not\preceq \mathbf{C}_{w,t}(x)) \vee (\forall \rho, N : (\langle \rho, N\rangle \in \mathit{Honesty}(x\ \mathsf{Intends}\ q,\ \mathbf{C}_{w,t}(x)) \wedge$ $\text{compatible}(f,\rho)) \rightarrow \mathit{Content}_f(x\,\mathsf{Intends}\,q) \not\supseteq \mathit{Content}_{f\circ\rho^{-1}}(N))] \Rightarrow$
  $(\exists \kappa, K : \langle \kappa, K\rangle \in \mathit{Honesty}(x\,\mathsf{Believes}\,\neg(x\,\mathsf{Intends}\,q), \mathbf{C}_{w,t}(x)) \wedge$ $\text{compatible}(f,\ \kappa)$ $\wedge \mathit{Content}_f(x\,\mathsf{Believes}\,\neg(x\,\mathsf{Intends}\,q)) \supseteq \mathit{Content}_{f\circ\kappa^{-1}}(K))$

Proof

Construct a canonical model as before. For brevity, let C refer to $\mathbf{C}_{w,t}(f_w(x))$. Then at any point $\langle w,t,f\rangle$, $\neg x\,\mathsf{Intends}\,p \in \langle w,t,f\rangle$ iff $x\,\mathsf{Intends}\,p \notin \langle w,t,f\rangle$, which is true iff $M \not\models_{w,t,f} x\,\mathsf{Intends}\,p$. But this is the case iff $\neg(\exists \rho, N : \langle\rho, N\rangle \in \mathit{Honesty}(x\ \mathsf{Intends}\,p, C) \wedge$ $\text{compatible}(f,\rho) \wedge \mathit{Content}_f(x\,\mathsf{Intends}\,p) \supseteq \mathit{Content}_{f\circ\rho^{-1}}(N))$. In turn that is the case iff either (1) $\neg(\exists \rho, N : \langle\rho,N\rangle \in \mathit{Honesty}(x\ \ \mathsf{Intends}\,p, C) \wedge$ $\text{compatible}(f,\rho))$; or (2) $(\forall \rho, N : (\langle\rho,N\rangle \in \mathit{Honesty}(x\ \mathsf{Intends}\ p, C) \wedge \text{compatible}(f,\rho)) \rightarrow \mathit{Content}_f(x\,\mathsf{Intends}\,p)$ $\not\supseteq \mathit{Content}_{f\circ\rho^{-1}}(N))$. Case (1) holds iff $x\,\mathsf{Intends}\,p \not\preceq C$. Thus the antecedent condition in the given constraint is met. Therefore, we trivially obtain $x\,\mathsf{Believes}\,\neg(x\,\mathsf{Intends}\,p) \in \langle w,t,f\rangle$ iff the constraint applies.

A more natural result is obtained in the presence of the constraint used in Theorem 4, which states that an agent can put parts of his cognitive state together.

Theorem 11 The addition of this axiom to Σ makes it sound and complete for models that satisfy the following constraints

- $[(x\,\mathsf{Intends}\,q \not\preceq \mathbf{C}_{w,t}(x)) \vee \neg(\exists \rho : (x\,\mathsf{Intends}\,q \not\preceq_\rho \mathbf{C}_{w,t}(x)) \wedge$ $\text{compatible}(f,\ \rho) \wedge$ $\mathit{Content}_f(x\,\mathsf{Intends}\,q) \supseteq \mathit{Content}_{f\circ\rho^{-1}}(\mathbf{C}_{w,t}(x)))] \Rightarrow$
  $(\exists \rho : \text{compatible}(f,\ \rho) \wedge x\,\mathsf{Believes}\,\neg(x\,\mathsf{Intends}\,q) \preceq_\rho \mathbf{C}_{w,t}(x) \wedge$
  $\mathit{Content}_f(x\,\mathsf{Believes}\,\neg(x\,\mathsf{Intends}\,q)) \supseteq \mathit{Content}_{f\circ\rho^{-1}}(\mathbf{C}_{w,t}(x)))$
- $(\bigwedge_{K\in \text{Cond}_{\mathbf{C}_{w,t}(x)}} K) \in \text{Cond}_{\mathbf{C}_{w,t}(x)}$ (assuming renaming of variables as needed)

Proof

For brevity, let C refer to $\mathbf{C}_{w,t}(f_w(x))$. Let L refer to the DRS $(\bigwedge_{K\in\text{Cond}_C} K) \in \text{Cond}_C$ (thus the variables in it are fixed). Thus every DRS, $N \in \text{Cond}_C$ is also in Cond_L, albeit with the variables in its universe renamed. Therefore, $(\forall \rho, N : (\langle\rho,N\rangle \in \mathit{Honesty}(x$ $\mathsf{Intends}\ \ q, C) \wedge$ $\text{compatible}(f,\rho)) \leftrightarrow (\exists \rho' : x\,\mathsf{Intends}\,q \preceq_{\rho'} L \wedge$ $\text{compatible}(f,\rho)))$— here ρ' is the obvious variation of ρ to account for the renaming of variables, when N is copied to obtain L. Now using the second constraint above (as well as the definition of ρ' above), we obtain the following: $(\forall \rho : \text{compatible}(f,\rho) \rightarrow (\exists \rho' :$ $\text{compatible}(f,\rho') \wedge (\forall N : x\,\mathsf{Intends}\,q \preceq_\rho N \wedge N \in \text{Cond}_C \wedge \mathit{Content}_f(x\,\mathsf{Intends}\,q) \not\supseteq$ $\mathit{Content}_{f\circ\rho^{-1}}(N)) \leftrightarrow \mathit{Content}_f(x\,\mathsf{Intends}\,q) \not\supseteq \mathit{Content}_{f\circ\rho'^{-1}}(L)))$. Now $\mathit{Content}_g(C) \subseteq$ $\mathit{Content}_g(L)$, because of the definition of $\mathit{Content}$. Also, if $\text{Domain}(g) \cap (U_C^* \cap U_L^*) = \emptyset$, then $Content_g(C) \supseteq Content_g(L)$. That is, under this condition , $Content_g(C) = Content_g(L)$. Since this condition holds for $f \circ \rho'^{-1}$, we have shown that, in the presence of the second constraint above, the antecedent of the first constraint is implied by the antecedent of the constraint of Theorem 10. Now we show that, in the presence of the second constraint above, the consequent of the constraint of Theorem 10 is implied by the consequent of the first constraint of this theorem. $K \preceq C$ implies that $K \preceq L$, for any K. Thus if $Honesty(K, C) \neq \emptyset$ then $(\exists \rho' : \langle \rho', L\rangle \in Honesty(K, C))$. We have already shown that, if $\mathrm{Domain}(g) \cap (U_C^* \cap U_L^*) = \emptyset$, $Content_g(C) = Content_g(L)$. Hence we have the desired result.

8. **Deliberate Non-Intentions:**

   The agent really intends not to have the intentions he does not have; i.e., the agent knows what he is opting out from.

   $\neg x\ \mathsf{Intends}\, p \rightarrow\ x\ \mathsf{Intends}\ \mathsf{A}(\mathsf{G}\neg(x\ \mathsf{Intends}\, p))$

   Theorem 12 The addition of this axiom to Σ makes it sound and complete for models that satisfy the following constraints

   - $[(x\ \mathsf{Intends}\, q \not\preceq \mathbf{C}_{w,t}(x)) \vee (\forall \rho, N : (\langle \rho, N\rangle \in Honesty(x\ \mathsf{Intends}\ q,\ \mathbf{C}_{w,t}(x)) \wedge \mathrm{compatible}(f, \rho)) \rightarrow Content_f(x\ \mathsf{Intends}\, q) \not\supseteq Content_{f\circ\rho^{-1}}(N))] \Rightarrow (\exists \kappa, K : \langle \kappa, K\rangle \in Honesty(x\ \mathsf{Intends}\ \mathsf{A}(\mathsf{G}\neg(x\ \mathsf{Intends}\, q)), \mathbf{C}_{w,t}(x)) \wedge \mathrm{compatible}(f, \kappa) \wedge Content_f(x\ \mathsf{Intends}\ \mathsf{A}(\mathsf{G}\neg(x\ \mathsf{Intends}\, q))) \supseteq Content_{f\circ\kappa^{-1}}(K))$

   Proof

   The proof in this case is similar to the one given above for negative introspection.

The above axioms all involved the agents' cognitive states at a given world and time, and expressed relations between parts of those cognitive states. It is also possible to state interesting and useful axioms in which objective facts about the relevant parts of the model can occur on one side of the $\rightarrow$. Some of these are enumerated below. In order to express these axioms, we need to extend the language with two more operators, $\Box$ and $\Diamond$, denoting truth at all worlds and times, and at some world and time, respectively. Formally,

- $M \models_{w,t,f} \Box K$ iff $(\forall w', t' : w' \in \mathbf{W} \wedge t' \in \mathbf{T} \rightarrow M \models_{w',t',f} K)$
- $M \models_{w,t,f} \Diamond K$ iff $(\exists w', t' : w' \in \mathbf{W} \wedge t' \in \mathbf{T} \rightarrow M \models_{w',t',f} K)$

9. **Consistency:**

   All intentions are potentially satisfiable; i.e., futile intentions are not held.

   $(x\ \mathsf{Intends}\, p) \wedge (x\ \mathsf{Intends}\, q) \rightarrow \Diamond((p \wedge \mathsf{P}q) \vee (\mathsf{P}p \wedge q))$

Theorem 13 The addition of this axiom to Σ makes it sound and complete for models that satisfy the following constraints

- $Content^I_f(\mathbf{C}_{w,t}(f_w(x))) \neq \emptyset$

Proof

Sufficiency: Let $x \,\mathsf{Intends}\, p$, $x \,\mathsf{Intends}\, q \in \langle w,t,f\rangle$. For brevity, let $\mathbf{C}_{w,t}(f_w(x))$ by referred to by C. Then $(\exists \kappa, K : \langle \kappa, K\rangle \in Honesty(x \,\mathsf{Intends}\, p, C) \wedge \text{compatible}(f,\kappa)$ $\wedge Content_f(x \,\mathsf{Intends}\, p) \supseteq Content_{f\circ\kappa^{-1}}(K))$ and $(\exists \mu, M : \langle \mu, M\rangle \in Honesty(x \,\mathsf{Intends}\, q,$

$C) \wedge \text{compatible}(f,\mu) \wedge Content_f(x \,\mathsf{Intends}\, q) \supseteq Content_{f\circ\mu^{-1}}(M))$. Define $\rho = \kappa \cup \mu$. Let $g = f \circ \rho^{-1}$. By Lemma 1 of §3.4, g is well-defined. By definition of *Content*, $Content^I_g(C) \subseteq (Content_g(K) \cap Content_g(M))$. Thus using the above constraint, we have that $(Content_g(K) \cap Content_g(M)) \neq \emptyset$. Using the above conditions (and Lemma 2 of §3.4), we obtain: $Content_f(x \,\mathsf{Intends}\, p) \cap Content_f(x \,\mathsf{Intends}\, q) \neq \emptyset$. Thus $(\exists w', t', S, t_1, t_2 : S \in \mathbf{S}_{w',t'} \wedge t_1, t_2 \in S \wedge M \models_{w,t_1,f} p \wedge M \models_{w,t_2,f} q)$. If $t_1 < t_2$, then we have $M \models_{w,t_2,f} (p \wedge \mathsf{P}q)$, else we have $M \models_{w,t_1,f} (q \wedge \mathsf{P}p)$. Thus clearly, $\Diamond((p \wedge \mathsf{P}q) \vee (\mathsf{P}p \wedge q))$.

Necessity: Using the definition of $Content^I$ and simple algebraic manipulations (as in the proof of Theorem 4), we can see that $M \models_{w,t,f} \Diamond((p \wedge \mathsf{P}q) \vee (\mathsf{P}p \wedge q))$ implies that $Content_f(x \,\mathsf{Intends}\, p) \cap Content_f(x \,\mathsf{Intends}\, q) \neq \emptyset$.

10. **Necessitation:**

    $\Box p \rightarrow x \,\mathsf{Intends}\, p$

Theorem 14 The addition of this axiom to Σ makes it sound and complete for models that satisfy the following constraints

- $Content_f(x \,\mathsf{Intends}\, p) = \mathcal{S} \Rightarrow (\exists \rho : x \,\mathsf{Intends}\, p \preceq_\rho \mathbf{C}_{w,t}(x)) \wedge \text{compatible}(f,\rho)$

Proof

For brevity, let C refer to $\mathbf{C}_{w,t}(f_w(x))$. As consequences of the definition of *Content*, we have the following (1) $(\forall N, f : Content_f(N) \subseteq \mathcal{S})$ and (2) for any f, for any w, t, $M \models_{w,t,f} \Box p$ iff $Content_f(x \,\mathsf{Intends}\, p) = \mathcal{S}$. For any point $\langle w,t,f\rangle$ in the model, $x \,\mathsf{Intends}\, p \in \langle w,t,f\rangle$ iff $M \models_{w,t,f} x \,\mathsf{Intends}\, p$, which is the case iff $(\exists \rho, N : \langle \rho, N\rangle \in Honesty(x \,\mathsf{Intends}\, p, C) \wedge \text{compatible}(f,\rho) \wedge Content_f(x \,\mathsf{Intends}\, p) \supseteq Content_{f\circ\rho^{-1}}(N))$. By (1) above and the definition of *Honesty*, this reduces to $(\exists \rho, N : N \in \text{Cond}_C$ $\wedge x \,\mathsf{Intends}\, p \preceq_\rho N \wedge \text{compatible}(f,\rho))$. But this condition holds iff $x \,\mathsf{Intends}\, p \preceq_\rho C$ $\wedge$ compatible(f,ρ). Hence the given axiom is determined by the class of models that meet the given constraint.

11. **Consequential Closure:**
$x \,\mathsf{Intends}\, p \wedge \Box(p \rightarrow q) \rightarrow x \,\mathsf{Intends}\, q$

Theorem 15 The addition of this axiom to Σ makes it sound and complete for models that satisfy the following constraints

- $[x \,\mathsf{Intends}\, q] \preceq_\pi K$ if $\Box(p \rightarrow q) \wedge [x \,\mathsf{Intends}\, p] \preceq_\pi K$

Proof

Let $x \,\mathsf{Intends}\, p \in \langle w, t, f \rangle$. For brevity, let C refer to $\mathbf{C}_{w,t}(f_w(x))$. Then $(\exists \kappa, K : \langle \kappa, K \rangle \in \mathit{Honesty}(x \,\mathsf{Intends}\, p, C) \wedge$ compatible$(f, \kappa) \wedge \mathit{Content}_f(x \,\mathsf{Intends}\, p) \supseteq \mathit{Content}_{f \circ \kappa^{-1}}(K))$. Given this, we have $K \in \mathit{Honesty}(x \,\mathsf{Intends}\, q, C)$ iff the given constraint holds. It can be easily seen that $\Box(p \rightarrow q)$ entails that $\mathit{Content}_f(x \,\mathsf{Intends}\, p) \subseteq \mathit{Content}_f(x \,\mathsf{Intends}\, q)$. Thus $x \,\mathsf{Intends}\, q \in \langle w, t, f \rangle$.

6 Conclusions

A formal theory is known by the inferences it keeps. We have considered several putative axioms that may be validated by imposing further constraints on the models in our theory. The basic logic does not validate several troublesome theses involving intentions and beliefs. The most important of those is closure under Logical Equivalence (we obtain the same results when beliefs are considered instead of intentions):

- $*$ $(x \,\mathsf{Intends}\, p) \wedge \Box(p \equiv q) \rightarrow x \,\mathsf{Intends}\, q$

Intuitively, this fails since x may not relate p and q in the appropriate manner: it is the structure of x's internal state that determines how he distinguishes among different conditions. In our theory, this inference can succeed only if q is subsumed by p, but in that case, the consequent is a fair characterization of x's internal state. Clearly that is not true for all such p and q. Thus this inference fails for the *right* reasons. This inference cannot be avoided in any possible worlds approach, not even those that consider "impossible worlds" [23], or "buddy worlds" [16] (roughly, because p and q are true at exactly the same worlds). One can, however, validate this axiom in our theory by adding the appropriate constraint (e.g., see Theorem 15 above).

While, by and large, this paper is in the spirit of Asher's paper on beliefs [5], it differs from it in some respects. Asher is more concerned with the philosophical issues involved in testing the correctness of belief reports; here we are interested in the logic of beliefs and intentions themselves and have, therefore, made some simplifications on grounds of technical clarity. The main conceptual distinction is in the way putative attitudes are evaluated—we consider a putative attitude by itself, while Asher considers the context in which an attitude is reported, including the cognitive state of the reporter. Two important technical differences are the following. Asher requires that the content of the subDRS of the cognitive state of an agent that matches the given belief report be non-empty (pp. 155, 173), i.e., the matching DRS be consistent. This can lead to non-monotonicity, a troublesome feature technically, in the following sense: a report that is correct with respect to a cognitive state can become incorrect as that cognitive state is extended. Asher also requires that "...the internal anchors of the subDRS generated by the report should approximate the contents of other components of the subject's total cognitive state which share reference markers with the belief." (p. 156) Thus if another part of the cognitive state than the one actually matched says that two reference markers be kept apart, they must be kept apart even in the one that is actually relevant. Thus Asher's account looks at the entire cognitive state. Another difference with Asher's paper is that we have actually enumerated several axioms and their corresponding constraints in this paper—this is not done by Asher, so this may be seen as a natural extension of his work.

About the only other formal theory of intentions is that of Cohen & Levesque [11]. This is a modal approach based on a possible worlds model. As a consequence, it validates closure under logical equivalence. It even validates a slight variation of closure under logical consequence, though an irrelevant reason prevents closure under standard logical consequence. This approach is very complicated, even though no additional axioms are considered: intentions are described as a third level concept, on top of other definitions. Thus it cannot easily be described and critiqued here.

We have presented some intuitions about intentions, and attempted to capture them in a formalism based on Discourse Representation Theory. We first characterized the minimal logic which yielded the most basic properties of intentions. We then presented a set of interesting axioms involving intentions, and the constraints required to validate them. These axioms would allow us to model agents who are more "intelligent" than the minimal agents considered initially or whose intentions are connected more tightly to their environment. This permits the application of our theory to a wide variety of AI problems. These axioms also may be used to motivate certain inferences that are invalid in general but may be acceptable in special circumstances, e.g., as models of how agents of a particular class deliberate, or used as heuristics or conjectures for reasoning in areas such as plan recognition [28]. In future work, we plan to incorporate an explicit account of actions and ability into this theory and also to extend it to the intentions of groups of agents. Another interesting set of problems, still to be addressed, concerns the complexity of the decision problems in the various logics that may arise from different combinations of the axioms that are considered in this paper.

References

[1] Philip Agre and David Chapman. Pengi: An implementation of a theory of activity. In *AAAI*, pages 268–272, 1987.

[2] James Allen and C. Raymond Perrault. Participating in dialogues: Understanding via plan deduction. In *Proceedings of CSCSI*, 1978.

[3] James F. Allen and C. Raymond Perrault. Analyzing intention in utterances. *Artificial Intelligence*, 15:143–178, 1980.

[4] Douglas Appelt. *Planning English Sentences.* Cambridge University Press, Cambridge, UK, 1986.

[5] Nicholas Asher. Belief in discourse representation theory. *Journal of Philosophical Logic*, 15:127–189, 1986.

[6] Nicholas Asher. A typology for attitude verbs and their anaphoric properties. *Linguistics and Philosophy*, 10:125–197, 1987.

[7] Nicholas Asher. Information, interpretation, and attitudes. In P. Hanson, editor, *British Columbia Studies in Cognitive Science*, volume 1. University of British Columbia Press, Vancouver, Canada, 1990.

[8] Michael E. Bratman. *Intention, Plans, and Practical Reason.* Harvard University Press, Cambridge, MA, 1987.

[9] Brian F. Chellas. *Modal Logic.* Cambridge University Press, New York, NY, 1980.

[10] Philip R. Cohen and Hector J. Levesque. Rational interaction as the basis for communication. Technical Report 433, SRI International, Menlo Park, CA, April 1988.

[11] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[12] Philip R. Cohen and C. Raymond Perrault. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3:117–212, 1979.

[13] Daniel C. Dennett. *The Intentional Stance.* MIT Press, Cambridge, MA, 1987.

[14] E. A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume B. North-Holland Publishing Company, Amsterdam, The Netherlands, 1990.

[15] Ronald Fagin and Joseph Y. Halpern. Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988.

[16] Ronald Fagin, Joseph Y. Halpern, and Moshe Y. Vardi. A nonstandard approach to the logical omniscience problem. In *Proceedings of the Third Conference on Theoretical Aspects of Reasoning About Knowledge.* Morgan Kaufmann Inc., 1990.

[17] Michael P. Georgeff. Planning. In J. F. Traub, editor, *Annual Review of Computer Science, Vol 2.* Annual Reviews Inc., Palo Alto, CA, 1987.

[18] Barbara Grosz and Candace Sidner. Attentions, intentions, and discourse structure. *Computational Linguistics*, 12(3):175–204, 1986.

[19] Hans Kamp. A theory of truth and semantic representation. In J. Groenendijk, T. Jansenn, and M. Stokhof, editors, *Truth, Interpretation and Information*, pages 1–41. Foris Publications, Dordrecht, The Netherlands, 1984.

[20] Hans Kamp. Context, thought and communication. *The Proceedings of the Aristotelian Society, New Series*, LXXXV(XIII):239–261, 1984/1985.

[21] Kurt Konolige. *A Deduction Model of Belief.* Morgan Kaufmann, Inc., 1986.

[22] Saul Kripke. A puzzle about belief. In A. Margalit, editor, *Meaning and Use*. Dordrecht Reidel, Dordrecht, The Netherlands, 1979.

[23] Hector Levesque. A logic of implicit and explicit belief. In *AAAI*, 1984.

[24] Diane J. Litman and James F. Allen. A plan recognition model for subdialogues in conversations. *Cognitive Science*, 11:163–200, 1987.

[25] Drew McDermott. A temporal logic for reasoning about processes and plans. *Cognitive Science*, 6(2):101–155, 1982.

[26] Barbara Hall Partee. Belief-sentences and the limits of semantics. In Stanley Peters and Esa Saarinen, editors, *Processes, Beliefs, and Questions*. D. Reidel Publishing Company, Dordrecht, The Netherlands, 1982.

[27] Raymond Perrault. An application of default logic to speech act theory. Technical Report 90, Center for the Study of Language and Information, Stanford, CA, March 1987.

[28] Martha E. Pollack. *Inferring Domain Plans in Question Answering.* PhD thesis, University of Pennsylvania, 1986.

[29] Stanley J. Rosenschein. Formal theories of knowledge in AI and robotics. *New Generation Computing*, 3(4), 1985.

[30] John R. Searle. *Intentionality: An essay in the Philosophy of Mind.* Cambridge University Press, Cambridge, UK, 1983.

[31] Robert C. Stalnaker. *Inquiry.* MIT Press, Cambridge, MA, 1984.