Requirements for belief models in cooperative dialogue*

Jasper Taylor and Jean Carletta
Human Communication Research Centre
2 Buccleuch Place
Edinburgh EH8 9LW
United Kingdom
jaspert@cogsci.ed.ac.uk

Chris Mellish
Department of Artificial Intelligence
80 South Bridge
Edinburgh EH1 1HN

March 2, 1995

This paper has not been submitted elsewhere in identical or similar form, nor will it be during the first three months after its submission to UMUAI.

Abstract

Models of rationality typically rely on underlying logics that allow simulated agents to entertain beliefs about one another to any depth of nesting. We argue that representations of individual deeply nested beliefs are in principle unnecessary for any cooperative dialogue. We examine some existing dialogue systems and conclude that where deeply-nested beliefs are supported, they are used only in situations where cooperation is not assumed.

Use of deeply-nested beliefs is associated with nested reasoning (i.e., reasoning about other agents' reasoning), which is limited by the architecture in most systems. Such systems cope well with cooperative dialogues because these do not require such reasoning. We show that summarizing representations of beliefs by conjoining all levels of nesting beyond the second allows agents capable of simple planning and plan recognition to cooperate in situations where agents equipped with unrestricted belief models would need to embody unrealistic assumptions.

*The work described here is funded by Joint Councils Initiative UK grant #SPG9200139, running for three years from November 1992.

1 Introduction

Most computational models of dialogue are based on a view of communication as plan-based rational action. Agents in these models have goals which they attempt to achieve by generating and executing appropriate plans. Utterances are seen as planned actions, chosen to achieve certain communicative effects. Planning is carried out in the context of a model of the state of mind of the planning agent, which is represented as a theory comprising a set of beliefs expressed in some logical formalism.

Since the effects of any utterance include changes in the state of mind of the addressee, a model of the speaker must also be capable of representing the speaker's beliefs about the mental state of the addressee. Typically, agents view their conversants as rational entities sharing their own capacities for beliefs and deductions but not necessarily their own actual beliefs, i.e., they have *nested beliefs.* The comprehension of utterances can be modelled as plan recognition; if an agent's plan refers to nested beliefs, then another agent's recognition of this plan must refer to the planner's beliefs about these beliefs, i.e., beliefs at one deeper level of nesting.

This paradigm suggests that an appropriate model for an agent engaged in dialogue should include a belief database capable of representing nested beliefs to any required degree. A number of languages of belief fulfilling this requirement have been developed from Hintikka's epistemic logic (1962), such as Appelt's (1985, Ch. 3) and Konolige's (1986).

However, in agent-modelling systems that provide such logics for agents reasoning about each other, deeply nested beliefs are seldom used. When they are used, it tends to be in the generation or recognition of an utterance that could be construed as involving some degree of deception. On the other hand, some systems embody limits on the depth to which their agents can hold nested beliefs, and these can model many aspects of cooperative dialogues. Furthermore, humans seem to have difficulty reasoning about individual deeply nested beliefs; there is evidence that young children handle them poorly, (Miller, Kessel, and Flavell 1970; Dunn 1991; Leekam 1991) and self-embedded structures of any kind (including nested beliefs in most logics) are generally the exception rather than the rule in human information processing. This sort of model seems too complex, appropriate only for deceptive or otherwise problematic dialogues.

We propose the following explanation for the apparent irrelevance of deeply-nested beliefs:

- There is a naturally distinguished subclass of dialogues, which we shall refer to as 'cooperative' dialogues, which are characterized by the lack of any commitment on the part of either participant to any form of deception, malicious or otherwise. Such dialogues are typical of task oriented domains and are the subject of most computer models.

- A property of such cooperative dialogues is that models of agents taking part in them need only distinguish between three types of belief according to nesting level.

Using such a restricted model of belief has the following additional advantages:

- Restricting an agent model to cooperative dialogue also limits the need for explicit inferences corresponding to deeply nested reasoning, e.g., reasoning about other agents' reasoning about other agents. Such inferences are not limited in standard belief logics.
- Where these restrictions result in differences in competence, that of the simplified model will be more human-like.
- We can choose the belief types to be used by the simplified model to be such as to allow reasoning with explicit representations of mutual beliefs. These, and other types of belief which can only be represented by infinite series in standard logics, have been postulated as fundamental entities by Grice, Searle and others. (Grice 1957; Searle 1969; Clark and Marshall 1981; Lewis 1969) Inferences involving them appear to be necessary for some types of communication to succeed.

We shall start in section 2 by looking at the history and current direction of the design of systems to simulate, or take part in, dialogue with humans. These embody a variety of relationships between logical principles and implementation details. In section 3 we look in particular at the support for nested beliefs in some of these systems, illustrating that while some systems allow, and make use of, private deeply-nested beliefs most have no need of them. In section 4 we look at the connection between the reasoning capabilities of dialogue systems and the nested beliefs they generate. We argue first that if cooperation is assumed there is no need to reason about other agents' recognition of plans. We show that nested beliefs are only required for nested reasoning, so cooperative dialogues only need beliefs at the object level and at two levels of nesting. Section 5 looks at the problems with such a model, particularly the fact that to simulate some communicative scenarios its agents are forced to adopt unsound or underspecified plans. These can be avoided by treating the last belief type in the representation as subsuming all deeply-nested beliefs, allowing them all to be inferred at once from certain situations to produce mutual belief. In section 6 we examine the performance of our model in the generation and recognition of referring expressions, showing that where its competence differs from what a system based on a fully-expressive belief logic could do, the former corresponds better to what humans tend to do. In section 7 we look at the likely capabilities and limitations of dialogue systems based on such models, and the possibility of needing to go beyond them.

2 The plan-based approach to dialogue

There is a broad-based movement regarding dialogue as purposeful, plan based activity, in the sense in which Miller, Galanter and Pribram (1960) hold plans to be the organizing principle of rational action in general. The early AI planners, such as STRIPS (Fikes and Nilsson 1971), used ends-means analysis to generate a set of actions that would transform an initial state of affairs in a physical domain into a state in which the planner's goals hold. When planning dialogues, the domain includes the beliefs and intentions of other agents, and of the planner itself.

States of belief are not inert like the block-stacking domains of the early AI planners; rather, they are subject to change as a result of reasoning by the believer. If a planner is to make use of language in a way that recognizes agents' capacity to reason, its model of the domain of beliefs must include a representation of the reasoning process. The paradigm for such models of rationality is formal logic, particularly Hintikka's epistemic logic (Hintikka 1962). There has been a clear trend towards the use of reasoning frameworks with the unrestricted power of theorem provers. At the same time, the planning process itself has also been evolving in this direction, with the situation calculus (McDermott 1982), modal truth criteria (Chapman 1987) and theory extension (Reichgelt and Shadbolt 1989) all providing means of equating planning with theorem proving in a logical problem space.

There now exists a range of implemented systems which either model dialogue between simulated agents or interact with a human user in something approaching natural language. These systems differ widely in the degrees to which they approach the accepted ideals in reasoning and belief modelling, in their use of facilities unrelated to the logical ideal, and of course in their capabilities. Most of them keep the beliefs and intentions constituting their agents' plans separately from their models of beliefs about the actual state of affairs, though these can be integrated. Different levels of expressiveness can be used for the same information according to context; KAMP (Appelt 1985) combined a fully-expressive logical description of his agents with a set of "action summaries" which could be used by a typical AI planning system. The unabridged description is used to check and debug plans. In general, implemented systems lag behind the theoretical models in the degree to which they approach ideals of rationality.

Subsection 2.1 will look at utterances as planned actions. There are a number of different ways in which utterances can be characterized as affecting beliefs, e.g, as 'sharing' or 'attempts to convince'. Subsection 2.2 will look at systems that understand dialogue by recognizing the plans that led to its generation. Subsection 2.3 will explore the moves that have recently been made in subsuming belief modelling, planning and plan recognition into a common theorem-proving framework. Finally, subsection 2.4 will examine the variations on this theme that have been proposed to avoid the problems of computational complexity and intractability, and to restrict the generation of simulated utterances which from a human point of view appear implausible because of the complexity of the reasoning involved in their generation.

2.1 Utterances as moves in a plan

When we engage in dialogue, we are planning to achieve goals which depend in some way on our conversants. These goals are not achieved directly by our utterances; in addition cooperation is necessary, provided by a helpful or trusting disposition or some formal authority relation between conversants. Communication plus cooperation results in plans being shared between conversants.

The actual effects following directly from utterances can be characterized in a number of ways. Searle (1969) asserts that an utterance conveys a desire on the part of the speaker for a certain change in the addressee's state, and is understood when the addressee recognizes a speaker's plan to produce understanding by that very recognition, thus bringing the two of them to a state of mutual knowledge. Many schemes, including implemented systems, have been based on this notion, but more recently there has been a move towards using reasoning to generate more of the attitude changes triggered by utterances.

2.1.1 Exchanges as moves in a joint plan

The generation of simulated dialogues within a planning paradigm was first done by Power (1974). In his system, the operations that constituted plans were modelled as being joint actions by the two agents; their occurrence depended on a well-formed initiation-response pair, such as a question and answer, having taken place. This design was contemporary with AI planners such as STRIPS which treated the world as an inert place in which the only changes that occur are those that are specified explicitly in the effects lists of the operations that are done, a restriction known as the STRIPS assumption (Waldinger 1977). With a planning system such as this, utterances must be characterized as joint actions if the effects of a reply are to depend on the initiation that went before. However, modelling the beliefs of conversants leads logically to modelling their reasoning, and a system that does this requires a planning component that goes beyond the STRIPS assumption.

Clark and Schaefer (1989), like Power, regard utterances as components of mutual acts, and hold that they only affect beliefs in the context of the complete discourse component of which they form a part. However, they do allow complete exchanges to be embedded within other exchanges, creating a treelike structure for the dialogue as a whole. Such a model works very well for dialogues that are allowed to run to completion, and is particularly good for handling misunderstandings and subsequent repairs, but assumes that the conversants bring each component to completion. In reality, some components are left incomplete, either due to inattention or possibly interruption, and agents' beliefs can still be affected by the utterances which formed these uncompleted components.

2.1.2 Reasoning about motivations

Cohen and Perrault (1979) cut out Searle's cycle of planning and recognition by characterizing utterances as establishing a belief on the part of the addressee that the speaker desires that a certain belief become mutual between the conversants. Such effects fulfill a precondition enabling a subsequent shared action of recognition, e.g., “convince” for assertions, to bring about the actual mutual belief to which they refer. While the omission of the mutual belief from the direct effects of utterances allows simple deceptions to be modelled, it does not cover double-bluffs, and the conditions for the subsequent “convince” action are ill-defined. However it does indicate the possibility of a characterization that works for all levels; after all, in the double-bluff case the speaker wants the addressee to believe that the speaker wants the addressee to believe the literal truth of the statement. This relates to Cohen and Perrault’s effects in the same way that they relate to effects based on assumptions of truthfulness.

The most comprehensive theories of utterance effects at the moment are those based on default logics, such as those described by Perrault (1990) and Appelt and Konolige (1988). CYNIC (Taylor 1994) is similar in that utterance effects depend on defeasible assumptions modelled in a reason maintenance system. In these theories, utterances have belief effects on the addressee that depend simply on their mood and content, but which are defeasible and can themselves be the subject of reasoning by the addressee which may give rise to beliefs regarding the speaker’s intentions. As the effects are related to individual utterances rather than dialogue exchanges, the model handles interruption, eavesdropping and other asymmetrical phenomena in a natural way. However, it is their property of allowing reasoning about the same utterance at different levels of abstraction that makes them important for this discussion; in a cooperative dialogue, each of the individual nested beliefs that make up mutual belief can be derived by reasoning at a different level.

2.2 Recognizing utterances as plan moves

If understanding an utterance is simply a matter of reconstructing the plan of which it formed a part, then some reasoning process embodying the reversal of the inferences required for planning should do the job. Such *plan recognition* has the problem that the observed moves may not uniquely define a plan. Kautz’ work on the general problem (Kautz 1987) made use of circumscription to rule out plans containing actions beyond those observed. Allen’s implemented system (Allen 1983), which used plan recognition to answer queries with extra useful information not specifically requested, used heuristics based on the likely plans of the questioner. A more general system for performing a similar task has been described by Appelt and Pollack (1990).

The many interpretations that can be placed on a single utterance correspond to recognition of different plans in which it could play a part. Searle (1975) described a theory of indirect speech acts using this principle, in which the possible relationships between an actual utterance and the intended goal to be recognized corresponded to the relationships between a plan’s goals and the moves making up that plan. The default theories mentioned in the last subsection attempt to provide formal models of this process. Of course, the number of pragmatic ‘parses’, or possible relationships between content and purpose, for an utterance in such a theory is vast. Similarly, for any particular communicative intention an agent might have, there exists a vast range of utterances which could conceivably lead to its satisfaction, given the right inferences on the part of the addressee.

Although there exist criteria for ruling out certain interpretations, such as Grice's (1957) felicity conditions and Brown and Levinson's (1987) criteria for politeness, there have been few attempts to formalise them. Grice (1975) suggests that the phenomenon of implicature is the result of such constraints; information can be communicated by an agent's choice of utterance, as well as by the effects of the utterance itself. This in turn allows the speaker to avoid utterances with unwanted implicatures, and hence theoretically the addressee can derive 'meta-implicatures' from such avoidance, and so on. A model of this process forms part of CYNIC (Taylor 1994, pp. 123–129).

2.3 Formalization of planning and plan recognition

The various attempts described above to formalize planning and plan recognition have met with success, but it must be taken into account that deduction, planning and plan recognition are in fact three different types of reasoning. Planning is a form of abduction (inferring the antecedents from the consequents) and plan recognition a form of induction (inferring a relationship between different observations). There are various ways of getting standard theorem provers to model such reasoning; the situation calculus allows a mechanical proof of the *existence* of a correct plan, but would need a metalanguage to describe its actual adoption. Kautz uses circumscription to elevate the inductive and inherently uncertain process of plan recognition to a matter of deductive certainty, but where the recognizer cannot observe the planner all the time, some uncertainty may be unavoidable. Allen's plan recognition system has two components, a bottom-up process which recognizes plans by induction from observed behaviour and a top-down, abductive process which generates possible plans for the subjects based on their likely top-level goals. CYNIC replaces the concept of belief with a set of three propositional attitudes, each supporting different inferences, allowing a single reasoner to combine the three reasoning modes.

However this is done, it opens up the scope for a large assortment of different types of nested reasoning. For a start, it is not just plans that can be recognized; so can material implications, deductions about them and other recognitions. Similarly, agents can plan for other agents to plan or do plan recognition, and indeed most combinations of types of reasoning can sometimes result in arguably plausible conclusions. The problem is that there are just too many possibilities; they grow faster than exponentially with the number of nested levels of reasoning.

2.4 Attempts to escape problems of theorem-proving models

Of course the problem of exponential growth of search space is a familiar one to AI. In the context of dialogue planning it is especially troublesome because of the phenomenon of Gricean implicature. In many domains requiring a solution to be chosen from an exponentially large space, finding something pretty good involves much less search than finding the optimal solution. However although a given intent can usually be served by any one of several utterances, using a non-optimal one may give rise to an unwanted implicature, with the addressee drawing false conclusions from the speaker's choice of that utterance rather than the optimal one. So any means of reducing the search space must be based on principles mutually understood by the conversants.

Human reasoning seems to differ from an unconstrained theorem-proving process in a number of ways, e.g., avoiding long chains of inferences and placing restrictions on which propositions can be combined to produce new conclusions — Fagin & Halpern (1988) describe some attempts to formalize these restrictions. People also have difficulty with repeatedly self-embedded structures, as illustrated by the example, "The rat the cat the dog chased bit died". The repeated centre-embedding in this sentence makes it hard to understand, although it is sanctioned by context-free models of sentence parsing.

Dialogue systems however have relied on models of reasoning which restrict the available inferences in less principled ways. Before the general acceptance of the theorem-proving nature of planning and plan recognition, these processes were implemented by custom reasoners frequently adapted from those used in single-agent robot problem-solving, like NOAH (Sacerdoti 1975). These made all sorts of approximations based on the typical shape of planning problems in inanimate domains, e.g., the STRIPS assumption. More recently, unrestricted reasoners have been adapted by adding explicit simplifications. A well-known example of these is KAMP (Appelt 1985); here, a fully expressive representation of the world is used for checking plans, but a set of simplified rules based on this model, called *action summaries*, are used for the planning process.

Other simplifications include the grouping of believed propositions into *topics*, and the representation of a set of beliefs about a topic as a *concept*. Ballim& Wilks (1991) describe a system for maintaining a set of nested beliefs by ascribing beliefs about topics to other agents, allowing deeply-nested beliefs to be computed as required. JAM (Carletta 1992) represents agents' acquaintances with concepts rather than all the believed propositions that constitute those acquaintances, which makes it easy for one agent to work out what another can actually do with their knowledge without necessarily sharing that knowledge.

2.5 Conclusion

When planning systems were first used for dialogue generation, they did not correspond very closely with the process of deduction in the logics that have been developed to describe the behaviour of beliefs; thus these systems had small repertoires of behaviour, and often very stilted conversational styles as a result of the utterance schemata being contrived to fit the capabilities of such planners. With the more recent generation of such systems, the flexibility of the theorem-prover approach allows the modelling of more complex and compact utterances (compare Power's "Shall we make a plan?" with Appelt's "Use the wrench in the tool box to remove the flywheel"). However this shift has brought with it a vastly expanded problem of search, and the need to identify the types of restrictions on inference that people actually embody themselves and expect others to embody when planning their utterances.

So far we have not discussed the nature of the restrictions on the nesting of beliefs embodied by existing systems; these will be discussed in the next section. Basically they are following the same pattern as the other restrictions; a shift from implicit, implementation-determined restrictions in systems based on the AI planning concepts, to explicit, logically neat restrictions placed on an underlying unrestricted theorem prover.

3 Nested beliefs in dialogue simulations

The paradigm underlying most recent AI dialogue simulations holds that the way to generate and recognize utterances optimally is to use a knowledge representation scheme that corresponds to a fully expressive logic of belief. Such logics are arrived at by replacing that axiom in a standard epistemic logic which constrains all knowledge to be true, with a weaker axiom that merely constrains each agent's beliefs to form a consistent set (Halpern and Moses 1985). They thus allow states in which agents believe their peers to be mistaken as well as simply ignorant. Fagin & Halpern (1988) have examined logics in which the axiom of consistency is weakened to avoid the problem of logical omniscience, all of which rule out some valid inferences their agents might otherwise make.

The implication is that the closer an agent model approaches an implementation of a logic of belief sanctioning logical omniscience, the better its performance will be in terms of the appropriateness of the utterances it generates and the correctness of its interpretation of incoming utterances, provided the reasoning component embodies the constraints that characterize well-formed utterances, e.g., avoiding incorrect implicature. But the logic of belief itself permits formulae expressing beliefs to any degree of nesting, and the rules of inference will actually generate infinite series of such beliefs from the most ordinary situations. For instance, if Fred and Doris catch one another's eye in an unexpected setting, they can each acquire:

- The *object level belief* that they are both present, which proposition we shall call P. These beliefs we will write DBP and FBP.
- The *first-level nested belief* that the other believes P. These we write FB-DBP and DBFBP.
- The *second-level nested belief* that the other believes P. These we write DBFBDBP and FBDBFBP.

- An infinite series of more deeply nested beliefs continuing the series.

The combination of all these beliefs is often called *mutual belief*, a topic we shall examine in section 5. The presence or absence of the individual, or private, deeply-nested beliefs can only affect the behaviour of the system if the planning or plan recognition component is capable of making use of them, i.e., if it is based on the theorem-proving approach to planning.

Clearly, representation and use of an unlimited space of nested beliefs inevitably leads to very steep combinatorial expansions when searching it – Carletta (1992, Ch. 6.5) illustrates that the growth is faster than exponential. But many dialogue systems embody some implicit limit on the depth of belief nestings, and do not appear to suffer any obvious loss of generality as a result. Power's (1979) model of two agents making joint plans to move about their domain includes private first-level nested beliefs, but does not appear to make use of them. For instance, an agent who hears another agent ask a question about some aspect of the world will record that his partner "cannot see" the aspect referred to. But when planning a question, no reference is made to whether or not the partner knows the answer; questions are represented in Power's planner simply as leading to knowledge of the answer, without reference to the process of reply. Indeed, these agents do not appear to use their nested beliefs at all; their plans can be private or mutual, and in the case of a mutual plan the agents are not concerned with the parts of the plan known only to the other agent. Hence agents know about operations in mutual plans, and know that the other agents know about them, but since this is the only context in which nested beliefs are used, this representation of *shared* knowledge is perhaps better regarded as one of *mutual* knowledge.

Early systems such as this tended to include customized reasoning processes which enabled them to display more sophisticated behaviour than their theoretical basis actually sanctioned. For instance, Power's example dialogue contains an instance of an agent expressing doubts about a plan and being reassured by his partner. It is not clear how John decides to express his belief that the MOVE operator achieves nothing, but the heuristic of accepting the most complicated available model assures that when Mary tells him that under certain circumstances the MOVE operator has effects, this becomes shared knowledge.

Operating in a similar domain, Houghton's (1986) system allows its agents to reason about other agents' mistaken beliefs about the domain, and therefore needs to represent nested beliefs. Houghton's system also has a significant component dedicated to grammatical choices, and hence uses nested beliefs to decide on the acceptability of various referring expressions, although this is done in an ad-hoc way. However there is no reasoning about other agents' use of reference, so no need for doubly-nested beliefs. Although the system is described as having mutual beliefs about which type of interaction its agents are engaged in, these are never represented explicitly, and in particular there is no scope for agents to be mistaken about interaction frames.

The success of these systems leads us to believe that some sort of limit on belief nesting levels has a place alongside other restrictions on inferencing that can be used in reasoning about dialogue within the theorem-proving paradigm. However it not very meaningful to compare the competences of different dialogue systems directly, as they address many different domains and styles of interaction. So we must instead look at the sort of problems that might require the handling of deeply-nested beliefs, and examine how systems with and without such beliefs address them. Such problems include:

- Interpretation of referring expressions,
- Deceptive or "political" dialogue,
- Utterances explicitly mentioning others' points of view.

In subsection 3.1 we will look at systems that support deeply-nested beliefs to see under which of the above circumstances they affect performance. In subsection 3.2 we will explore the limitations on systems that don't make use of deeply-nested beliefs, and look at the assumptions they employ when a situation would seem to require this capability. We conclude that situations in which deeply-nested beliefs can make a difference are always characterized by deception.

3.1 Systems with deep nestings

Here we describe systems that attempt to capture the unlimited belief nestings sanctioned by standard logics of belief. The first is a model of belief ascription, the second a system for teaching concepts, and the last an utterance planner/recognizer.

3.1.1 ViewGen

ViewGen (Ballim and Wilks 1991) is a model of an agent capable of individuating beliefs at different levels of nesting to arbitrary depths. It is intended to illustrate a set of heuristics for generating deeply-nested beliefs when these are not directly available to the agent. Although it does not include an utterance planner or recognizer it is intended as a component for a simulated agent, and the authors provide many examples of possible dialogues between the system and other agents, along with the beliefs that play a role in them.

The first of these runs as follows (p. 32):

Example 1
USER: Frank is coming tomorrow, I think.
SYSTEM: Perhaps I should leave.
USER: Why?
SYSTEM: Coming from you that is a warning.
USER: Does Frank dislike you?
SYSTEM: I don't know, but you think he does and that is what is important now.

The last line of this is an example of an utterance making explicit reference to nested beliefs, but the interesting point here is the derivation of SYSTEM's first utterance, for which this is generated as an explanation. USER's initial utterance has been taken as a hint, based on what SYSTEM thinks USER thinks about Frank's attitude to SYSTEM. Writing P for the proposition that Frank does not like SYSTEM, the condition the author gives for this interpretation is that SBUBP (SYSTEM believes that USER believes that P). Actually this is not sufficient: if in addition SBUBSB(not P), the hint would not work. Even given the absence of this belief (in which case its converse can be added by the ascription mechanisms described), the 'hint' interpretation appears to depend on SYSTEM believing that USER has some good reason for not simply saying, e.g., "Frank is coming, so you had better leave". The imperative to be polite is such a reason, but "Frank is coming tomorrow, I think." only fulfills Brown & Levinson's (1987) requirement of not constraining the addressee's options if SBUB(not P). Thus in picking up the hint, SYSTEM is interpreting USER's utterance as deceptive, inasmuch as it conveys UBSBUB(not P).

This is the only example the authors give of deeply nested beliefs in interpretation of the intent behind utterances. Subsequently they present some examples of the generation of utterances appropriate to the addressee's level of knowledge (about the inherited disorder thalassemia), but these only require one level of belief nesting. The salient point about the SYSTEM/USER/Frank example is that it requires deep-nested beliefs, but the context involves recognition of a deceptive stance; the explicit mention of a deep-nested belief comes later.

3.1.2 Kobsa's BGP-MS

BGP-MS (Kobsa 1990) stands for Belief, Goal and Plan Maintenance System. It is a user-modelling component which has been incorporated into a natural language interface to expert systems, and Kobsa's examples are drawn from user interactions with an automated financial advisor. Its database comprises both explicitly propositional and conceptual information expressed in a semantic network-like structure, and like ViewGen it allows heuristics to ascribe beliefs to users on the basis of stereotyping. This system also has the potential to discriminate between individual nested beliefs down to any depth of nesting, while allowing deeply-nested beliefs at different levels to be conflated if they do not need to be individuated. In addition to SB for "the system believes that" and UB for "the user believes that...", Kobsa writes MB for "it is mutually believed by the user and the system that...", allowing combinations such as SBMBUBX: "the system believes that it is mutually believed by the user and the system that the user has concept X". Since all beliefs modelled are beliefs of the system, all belief labels begin with SB. Concepts in the system can be placed in different partitions according to what beliefs the system is representing about them. To find the state of belief about a concept, the tree of partitions is searched starting with the top level, SBMB.

A partition with an MB in its label can be split into a series of subpartitions by substituting the MB with all possible strings composed of SB and UB, including the empty string, in which the complete label does not contain a pair of adjacent SBs or UBs (these are equivalent to single SBs or UBs given an S5-based belief logic). Thus the partition SBMB can be split into SB, SBUB, SBUBSB, SBUBSBUB, etc. and these reassembled into subpartitions in a number of ways.

If a concept is in the SBMB partition, then it is believed by the system to be the subject of mutual acquaintance, but if not it may be in one or more of its "primitive" subpartitions SB, SBMBUB and SBUBMBSB, all three of which together are equivalent to SBMB. These partitions can be further divided, for instance SBMBUB divides into SBUB and SBUBMBUB.

BGP-MS can tailor the utterances it generates to take into account the user's level of familiarity with different aspects of a wide range of topics. However, Kobsa mentions that the system never actually generates nested beliefs beyond the fifth level, e.g., SBUBSBUBSBUBX, and the examples he gives do not even require creating a distinction between third-level nested beliefs and deeper-nested ones. He then goes on to say that in the general case of cooperative person-computer dialogues the partitions representing private nested beliefs (those not referring to mutual belief, i.e., SBUB, SBUBSB, etc.) are empty. He claims that such beliefs do occur when modelling uncooperative situations such as those in which deception must be considered, but does not discuss any examples of BGP-MS being used in such situations.

3.1.3 Taylor's CYNIC

CYNIC (Taylor 1994) models the generation and interpretation of utterances by agents who do not necessarily cooperate. Its database is somewhat restricted in respect of the complexity of the domains with which it can cope. It has unique names for all individuals, can handle only explicitly propositional information and has no representation for any kinds of mutual belief, except in that plan operation characteristics and rules of inference are implicitly mutually known. What it does have, though, is a reasoning system equivalent to a theorem prover in a logic encompassing desires as well as beliefs, and a set of rules for deciding when utterances are inappropriate due to impoliteness or incorrect implicature. When generating the indirect request, "Can you pass the salt?" it is actually planning to have the addressee demonstrate their ability to pass the salt by doing so, a process requiring third-level nested attitudes in the planner.

Writing S (Speaker), A (Addressee), B (Believes), W (Wants), C (A can pass the salt), P (A passes the salt), I (A informs S that C) and Q (S asks A whether C) the speaker's reasoning goes like this:
SWP (premise: speaker wants the salt passed)
SWAWP (for which addressee must want to pass it)
SWAWSBP (which he will if he wants her to believe it is passed)
SWAWSBC (or that he can pass it.)
SWABSWSBC (So she plans for him to adopt her goal of knowing he can pass it)

SWABSWSBI (which he is to deduce from her goal of being told he can)
SWABSWI (i.e., of him telling her he can)
SWABSWAWI (or him wanting to tell her)
SWABSWABQ (which comes from her goal of him having been asked)
SWABSWQ (i.e., of her asking him)
SWABQ (which he would know was her goal, because she did it)
SWQ (so she wants to do it)
Q (so she does it.)

Examples involving attempts at deception go even deeper; the system can both generate and recognize the double-bluff in the following exchange between two spies:

Example 2
Sergei: "The secret weapon is in Dneprpetrovsk."
Natasha: "Ah, I guess the secret weapon is *in Dneprpetrovsk. You wouldn't say that unless you wanted me to think it was in Krasnoyarsk!"*

Sergei's utterance takes four nesting levels to generate and five to recognize correctly; where D is the proposition that the weapon is in Dneprpetrovsk, this recognition requires that NBSWNBSWNBD. However for more straightforward communicative acts, the control heuristics ensure that plans involving only shallow-nested beliefs are preferred; deep nestings such as these are only produced when shallow-nested plans are blocked, specifically by politeness axioms or by removing the assumption of helpfulness.

While the deeply-nested beliefs in the deception examples seem plausible, the reasoning illustrated above for an indirect request seems overly complex as, unlike the "hint" in the example from ViewGen, no deception is involved. The complexity appears to stem from the system's inability to use mutual beliefs, whose consequences in indirect requests we will return to in section 4.4.1.

3.2 Dialogue simulators without deeply nested beliefs

We now look at a dialogue system with an explicit belief nesting limit, and a planning assistant with an implicit nesting limit.

3.2.1 Carletta's JAM

JAM (Carletta 1992) is a dialogue simulation that works in the domain of the *map task*, a simple navigation problem where one person (the instruction giver or IG, typically "she") guides another (the instruction follower or IF, typically "he") along a route by referring to landmarks they might encounter along the way. The task is complicated by the possibility that not all the landmarks appear on both maps (see Figure 1 for an example). JAM is able to produce simulated dialogues that appear to capture the essence of many of the interactions that are observed in the corpus of human performances (Anderson, Bader, Bard, Boyle, Doherty, Garrod, Isard, Kowtko, McAllister, Miller, Sotillo, Thompson,

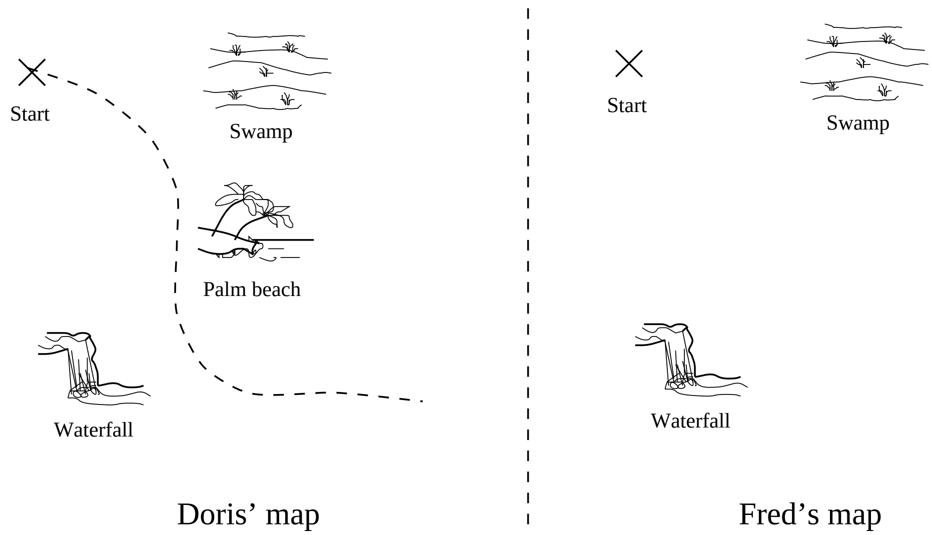

Figure 1: Part of a typical pair of maps in the Map Task

and Weinert 1991) in this task. By changing the capabilities and preferences of the planning component of the system, different styles of dialogue can be generated.

JAM's belief model represents agents' knowledge of the maps in terms of degrees of acquaintance with concepts corresponding to the landmarks and their interrelations. It limits beliefs to object level plus singly- and doubly-nested, and for the latter it conflates uncertainty about beliefs at different nesting levels (e.g., "He is uncertain about my belief" and "He believes I am uncertain"). Each agent summarizes the point which discussion about each concept has reached by representing it as being in one of 18 possible states, each state corresponding to a combination of object level, singly- and doubly-nested acquaintance.

An example for the maps illustrated might go as follows:

Example 3
Doris: "The first section of the route goes left of the palm beach.
Fred: "I don't understand. Where is the palm beach?"
Doris: "The palm beach is below the swamp."
Fred: "OK. I am done talking about the first section."

In this example, Fred's question in line 2 is a result of his recognizing and adopting Doris' plan for him to become acquainted with the first section, and detecting the lack of the condition FBDBFB(Swamp), where (Swamp) is the information needed to locate the swamp. His utterance "I am done..." in line 4 results from his prior belief that he is fully acquainted with the first section but is uncertain about Doris's belief that he is so acquainted, i.e., FBDBFB(Section 1).

We have enhanced JAM by adding a belief model with a variable limit on nesting and no conflation (Taylor 1993). From this we have found that reducing the nesting limit to allow only first-level nested beliefs results in simpler but longer dialogues, for instance preventing the short cut embodied by the utterance of "I am done..." in the above example. However, increasing the nesting limit does not result in any difference in performance from the original version. This is despite the fact that JAM's planner contains operations with second-level nested beliefs as effects, and plan recognition involves reversing the planning process at one extra level of nesting, and thus could require third-level nested beliefs.

The agents in JAM can be set up to cooperate actively in various ways including recognizing and adopting one another's goals, or not to cooperate beyond making appropriate responses when addressed. However, the task-oriented character of the domain rules out any deception or private goals, and in the examples handled by JAM there is not even the possibility of agents having mistaken beliefs. So it is possible that JAM's good coverage of interaction types from the Map Task corpus is related to this feature of the task. In section 6.1 we will argue that adding ambiguity to this task domain does not require deepening the belief nesting limit.

3.2.2 Allen et al's TRAINS

The TRAINS project (Allen, Schubert, Ferguson, Heeman, Hwang, Kato, Light, Martin, Miller, Poesio, and Traum 1994) is a large-scale demonstrator project for natural-language interface technology in the realm of communication between human users and automated expert assistants. The domain chosen for the development is a simplified industrial world, in which trains move materials and products between factories and warehouses. The project considers the user and system to be "peer agents", combining the human's flair for strategic thinking with the machine's ability to check details thoroughly and spot possible conflicts. An important characteristic of the system is the complexity of the domain; it involves simultaneous actions, actions with durations and actions whose effects depend on the state of the world at the time they are executed. Such domains require a fairly up-to-date AI planner even before the extra requirements of dialogue generation and understanding are considered.

The current version, TRAINS-93, contains no explicit representation of beliefs about the state of the domain, which is modelled as being directly accessible to both the user and the assistant so their mutual beliefs about it are identical to the facts. However there are different types of attitude to the plan operations under discussion. The system places operations comprising the plan under development into various categories according to the stage reached in discussing them. The system's proposals start out as part of its private plan, and when first mentioned in the dialogue they become part of its proposed plan. The user may acknowledge the suggestions, causing the corresponding plan operations to become shared in their scope as proposals. Only when the user explicitly accepts them do they become part of the shared plan. The system also models

TRAINS category	Belief type
System's private plan	SB
System's proposed plan	SBMBSB
Shared plan	SBMB
User's proposed plan	SBMBUB
User's private plan	(unknown to system)

Table 1: Nesting levels coresponding to TRAINS belief types

the movement of operations proposed by the user from their proposal to their inclusion in the shared plan by similar criteria.

These categories correspond to states of belief about the ultimate execution of the operations, from private, via private nested and shared to full mutual belief. In terms of Kobsa's notation (introduced in section 3.1.2) the types of belief in TRAINS correspond to combinations of nested beliefs as shown in table 1.

The updating of beliefs towards shared status corresponds only roughly to inferences in standard belief logic, and it is not clear what happens when an operation previously in the plan has to be removed. Inferences about the effects of the plan are made using event-based temporal logic, whereas referring expressions in utterances relating to it are resolved by processing a separate representation in episodic logic, a language more structurally related to English.

Essentially, this is another example of a system designed exclusively for co-operative interactions and making no distinctions between beliefs with status differing only beyond the second level of nesting, i.e., the same cutoff level as JAM. [1] In both cases, utterances result in beliefs being transferred between members of a predefined set of states. Here, though, the cutoff seems to be a result of the design of these transitions, whereas in JAM the states were defined by starting with a belief model discriminating between all possible nesting combinations, cutting away that which appeared superfluous and then encoding the logical deductions that remained in terms of transition networks.

3.3 Conclusion

We have seen that typically, dialogue simulations do not model nested beliefs beyond the second level. Those that do so either make little or no use of the capability, or are engaged in the generation or recognition of utterances that in some way serve the agenda of one agent to the exclusion of the other. Even systems capable of discussing actions in complex domains, and of generating and interpreting referring expressions, do not go beyond doubly-nested beliefs, though we shall return to the discussion of referring expressions in section 6

[1] in fact, proposed plan entities are further divided into private and shared, depending on whether they have been acknowledged or not, but it appears that no reasoning is done with the subjects of unacknowledged proposals.

as it could be argued that some notations for the objects of references conceal representations of deeply-nested beliefs.

For the most part, dialogue systems are built around the assumption that the two agents being modelled, or the simulated agent and the human user, are engaged in cooperative problem-solving. Apart from this characteristic, the limitation of required nesting levels appears to have nothing to do with the actual capabilities of the agents, or the complexity of the domain.

There does not seem to be any reason why systems with limited belief nesting levels should not be set up to benefit from, and attempt, deception; indeed only a single level of nesting is required to differentiate between what one believes oneself and what one wants the other guy to believe. The double-bluff example from CYNIC, though, leads us to believe that our initial assumption that the more nesting levels a system can manage the better it will perform holds true when deception is under consideration.

4 Nested beliefs and nested reasoning

We shall now attempt to find a way to explain the patterns of nested belief usage we have noted. We have seen that in JAM, as in Power's and Houghton's systems, cooperation between agents is assumed throughout, and that these systems do not use third-level nested beliefs at all. In CYNIC, agents need not always cooperate (the "can you pass the salt" example was generated by blocking the assumption that a direct request would receive cooperation), and deeply-nested attitudes result. Kobsa has found that while BGP-MS has the capability to represent private deeply-nested beliefs, it is seldom used, and puts this fact down to the combination of co-presence and cooperation between system and user in his examples. So how is it that cooperation reduces the need for nested beliefs?

4.1 Utterance effects that come before cooperation

In order to describe the process of reasoning about dialogue, we must first decide on a description of the logical properties of the individual utterances themselves. For the reasons given in section 2.1.1, we must look to the models of dialogue that specify effects for individual utterances, and have some of these effects result in the emergence of a dialogue characterized by turn-taking.

To build a model that can encompass reasoning in a range of deceptive situations as well as honest ones, we consider the utterance effects that are *available given trust* in any situation, for instance that an agent making a declarative utterance believes the proposition expressed and is correct in this belief. These are not reliable, but they are invariably the basis on which reasoning about utterances proceeds; if you don't believe what someone tells you, you still think they were trying to get you to believe it, or to give you the impression they were trying thus, or to make some other deduction based on the utterance in a similar way. Many linguistic phenomena, for instance indirect speech acts and

Utterance mood	Addressee's response
Indicative	Believe proposition expressed
Interrogative	Intend to give literal answer
Imperative	Intend to behave obediently

Table 2: 'Canonical' effects of utterances

implicature, can be accounted for in terms of this type of circumspective reasoning about basically simple utterance effects; see Searle (1975), Grice (1975) and others. Implemented systems that make use of such defeasible inferences for utterance comprehension are described in section 2.1.2.

We thus require each utterance to have two effects; one that simply permits the continuation of the turn-taking structure of the dialogue, and another specifying the defeasible inferences made available by the utterance's content. The latter, or 'canonical', effects are shown in table 2.

4.2 Only nested reasoning requires nested beliefs

If an agent model simply reacts to the utterances addressed to it in accordance with these simple effects, i.e., it answers questions and responds to declaratives according to its belief model, it is readily apparent that its responses will generally depend solely on beliefs at the object level. The exception is the case where the propositional content of the utterance is itself a belief, e.g., "Does your mother know you're here?"; this will be dealt with in section 4.3.1. Normally, nested beliefs will make no difference, so there is no reason to distinguish between private beliefs and mutual beliefs in such an agent model. To react to an assertion that P, an addressee need merely believe it: ABP.

However, planning in any domain involves having beliefs about it, so a simulated agent that plans and executes utterances assuming such purely reactive behaviour on the part of its conversant must consider the effects of utterances in terms of the belief state of the addressee, and distinguish these from its own beliefs relating to the same propositions, but not necessarily from more deeply nested beliefs; (stating that P) results in SBABP, even if not SBP.

Plan recognition, modelled as inductive inference, involves fitting the actions an agent is observed performing into a plan which is then ascribed to them. To recognize a plan that has been generated assuming reactive behaviour, i.e., characterized by first-level nested beliefs, a simulated agent needs to be able to distinguish nested beliefs up to the second level. A's recognition (of S's plan (to state that P)) may result in ABSBABP, even if not ABSBP or ABP.

Now as we pointed out in section 2.3, if our system can deduce, plan and recognize plans, it may also be expected to deduce plans, plan for deduction, recognize deduction, and so forth. If agents are modelled as being capable of such nested reasoning, and also model each other as being similarly capable, they will be able to nest their reasoning more deeply than this, for instance making plans involving (possibly incorrect) plan recognition by others. However, most imple-

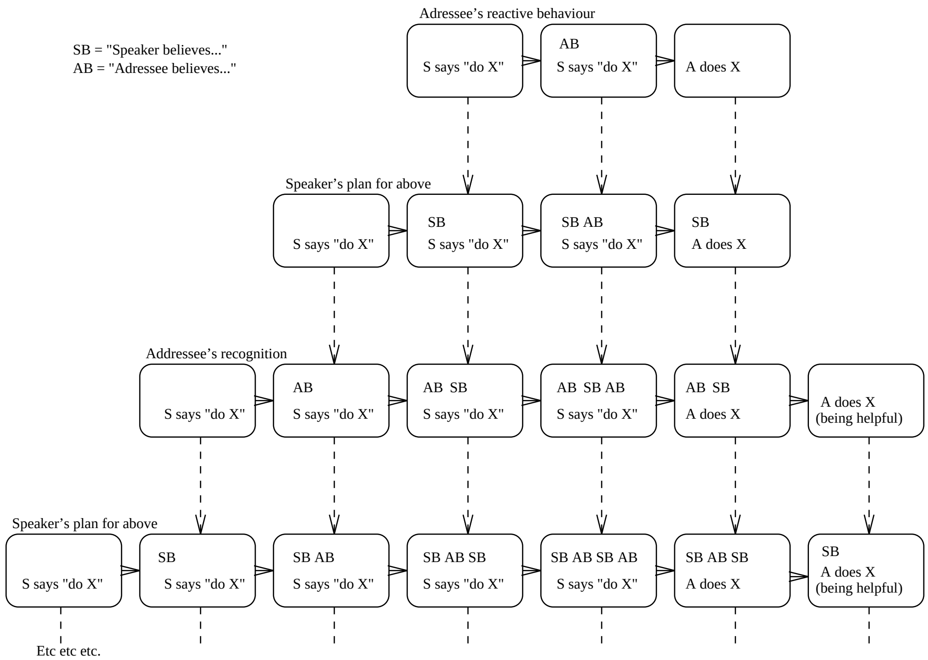

Figure 2: How each level of reasoning models a simpler one

mented systems combining planning and plan recognition components cannot make such plans. This is typically because the operations from which they can build plans do not include the act of plan recognition, and hence such an act cannot play a role in any plan another agent may generate.

Systems that are capable of modelling the generation of such plans will include beliefs and intentions specifically at the third level of nesting, and further iterations of nested reasoning will require still more deeply nested beliefs. Figure 2 illustrates how these deeply-nested beliefs are related to each other. A reactive response (the top layer) serves as an agent model in a straightforward plan for communication (the second layer), which in turn serves as a model in the recognition of the plan (the third layer). Planning for that recognition in turn requires another level beyond the level at which the recognition is intended to take place, and further abstractions of the same type are possible.

4.3 Cooperative dialogue needs plan recognition but nothing deeper

If two agents are conducting a dialogue in pursuit of shared objectives, e.g., a task-oriented dialogue, then they are attempting not merely to convey information to each other but to come to a state of mutual knowledge about this information. Searle claims that this is achieved when an addressee understands the meaning of an utterance by recognizing the speaker's intention to produce that understanding (Searle 1969). This sounds circular at first, but it can be interpreted as claiming that a speaker speaks in accordance with her plan and *intends that plan to be recognized, but does not plan the steps that constitute its recognition.*

To put it another way, the speaker makes a plan involving reactive behaviour by the addressee, and then executes it not in the hope that it will succeed as conceived (as the purely reactive characterization is rather too simple in most cases) but as the best way of getting it recognized, which results in success if combined with assumptions of helpfulness and honesty. Explicit planning for plan recognition (hereinafter PFPR) is redundant in such situations, as are deeper types of nested reasoning, so only object-level beliefs and nested beliefs at the first and second levels are used.

So here we have an argument which says quite simply that if two agents are cooperating with each other, they never need third-level nested beliefs. Each agent is quite content for the other to recognize her plans as they actually are, and therefore does not need to keep track of differences between her actual plan and what her conversant thinks her plan is. Plan recognition itself is useful for efficient cooperation, as it allows the recognizer to take the initiative in achieving the planner's goals, so a cooperative reasoner needs at least second-level nested beliefs. The definition of cooperation here does not rule out all deception; only that to which a conversant is *committed.* A cooperating conversant may say something which leads to her addressee forming a belief the speaker believes is false. The speaker could simply fail to foresee this, or anticipate it (if the false belief is the result of deduction rather than plan recognition) but not think it important. The case of oversimplifying a subject for educational purposes is also cooperative; we will return to this in section 4.4.2.

4.3.1 Utterances explicitly about beliefs

It is not too hard to imagine cases of cooperative dialogues that do need deeply-nested belief; however, given cooperation it seems there are always simpler ways of conveying the same information that do not. Suppose one person is *honestly* relating a chain of events to another person, and in that chain of events a third person relates another chain of events, and in that.... If any of the stories within the current story involve deception, they require deeper nesting for their comprehension than that exhibited by the characters involved in them. It makes little sense to include such narratives within our definition of cooperative dialogue, since the listener must be capable of following whatever mental processes the deceitful characters are deploying.

On the other hand, if all the stories are honest, it is not obvious that there really is any need to narrate anything beyond what actually happened and who is in the know about it; the exact path by which the news reached the story-teller, and the details of who knows who else knows, are generally unimportant, so the whole process can be handled within the competence of the belief model required for cooperative dialogues that don't involve third parties. These admittedly foggy intuitions will be fleshed out later in section 6, which describes what happens when cooperative agents get confused about one another's use of referring expressions.

4.4 Deception requires PFPR, but politeness and deliberate simplification do not

There are situations in which a speaker does need to plan for the recognition of a particular plan by her addressee. Attempted deception is the most obvious, where the plan to be recognized contradicts the actual plan. However such deeply nested reasoning also appears to characterize some utterance forms used in the service of politeness; a speaker suggesting his partner do some action may not use the direct request that would constitute a plan for reactive behaviour, but might say, e.g., "It's cold in here" in order to have a plan recognized in which the addressee attempts to make it warmer by, say, closing a door. In this example, it appears that the addressee does indeed recognize the speaker's plan that she close the door, rather than just recognizing a plan to communicate the fact that it is cold and then helpfully making it warmer.

4.4.1 Statements as indirect requests

In the example in subsection 3.1.3, an agent utters "Can you pass the salt?" intending the addressee to run through the possible responses to the request and conclude that to demonstrate her ability to pass it by doing so is the simplest which might have useful consequences for the speaker. This example involves third-level nested beliefs on the part of the speaker, who is quite clearly planning for a particular plan to be recognized. However, the necessity of such planning is merely a consequence of the system's preference for plans with a direct causal chain between the operations and the outcome. A more traditional Searlian view of such a request holds that the speaker, unable to make a direct request, is establishing a condition for the passing of the salt, after which he may ask for it to be passed. By this interpretation, the deepest nested attitude on the speaker's part is SBABC, though, like other plans for reactive behaviour, this is intended to be recognized rather than to succeed as such. In either case, the addressee need only carry out straightforward plan recognition to decide that Fred's intention is that she pass the salt; ABSBABC implies ABSBABP.[2]

In a similar manner we can model the case of Fred uttering "It's cold in here" as a request to Doris to shut the door. The recognition process is uncontroversial, and consists of Doris arriving at DBFBDBC (C stands for the proposition "It is cold") and hence DBFBDBS ("Doris will shut the door"). Fred can plan this either by reasoning exactly as Doris recognizes, in which case he genuinely intends her to believe that it is cold, or orchestrate Doris' response by PFPR, in which case he needs the third-level nested belief FBDBFBDBS. It seems unlikely that Fred should try to convince Doris of something her own skin is already telling her the truth or falsity of, so is PFPR really taking place?

[2]It might also be argued that such utterances as this have become synonymous with direct requests, as evidenced by the fact that they can be augmented with "please".

We would say yes, but it is actually concurrent with the exercise in straightforward planning in which Fred plans for Doris to come to believe his statement that it is cold, then make her own plan to rectify the situation. Like the plan to find out if someone can pass the salt, the beliefs in this plan may not be actually held. However, in the case where Fred thinks it is *not* cold, but wants the door shut for some other reason, then he must be doing actual PFPR, but in this case we are back to deception.

4.4.2 Simplifications for teaching

Another interesting half-way point, of relevance to user modelling, is use of a deliberately simplified scenario in teaching. Physics teachers describe Newtonian mechanics even when they know about Einstein's general relativity, and programming can be taught using a "virtual machine" without mentioning that the actual machine is engaged in a lot of other tasks at the same time. We categorize such situations as cooperative because the teacher has no commitment to the deception. If the student suspects she is not being taught the whole truth, the teacher can come clean and proceed with the lesson with the mutual knowledge that what is being taught is a simplification.

If a simplified lesson is to be any use at all, the concepts as described in the lesson (e.g., mass, velocity, etc. in the case of physics) must relate to each other *approximately* the way they do according to the teacher's beliefs. This will involve some relationships between concepts being described in terms of propositions which are believed false by the teacher. In the normal case of such a lesson, the beliefs of both teacher and student are exactly what they would be in a straightforward instructional dialogue, with the exception of the teacher's lack of belief in the content of the lesson. For instance, suppose the teacher T tells the student S proposition P; that force equals mass times acceleration, which is true in Newtonian physics but not in general relativity. The intended result is that TB(not P), TBSBP, TBSBTBP, SBP, SBTBP, SBTBSBP. If the student knows P is false, it is no problem to keep quiet about it, in which case the resulting beliefs are TB(not P), TBSBP, TBSBTBP, SB(not P), SBTB(not P), SBTBSBP.

If the teacher knows the student knows enough about general relativity not to believe P when told so, then professing P might be expected to lead to TB(not P), TBSB(not P), TBSBTB(not P), TBSBTBSBP. i.e., P is believed only at the third level of nesting. The teacher would only attempt this as part of a plan to be recognized as wanting to teach P. This is not compatible with cooperative interaction; it seems natural instead to avoid the complication by teaching P', the proposition that force is almost exactly mass times acceleration except at speeds approaching that of light. P' is a belief that can be shared by teacher and student at all nesting levels.

4.5 Conclusion

Nested beliefs go hand in hand with nested reasoning, i.e., reasoning about what inferences another agent is likely to make. While second-level nested reasoning, e.g., plan recognition, is helpful to the pursuit of cooperative goals, nothing beyond this ever is. This is because plan recognition allows the addressee to realize when the speakerer is mistaken about the domain or the addressee's beliefs about it, but anything at a deeper level, e.g., PFPR, would only need to be pursued if the reasoning contained within it two levels shallower (e.g., the plan to be recognized) did not actually happen, i.e., for deception.

Hence, in a world in which deception is unknown, agents never find such nested reasoning necessary, and thus need not represent the nested beliefs at the third level and beyond, which it manipulates. Furthermore, the kind of deception that does require third-level nested beliefs involves agents presenting false models of themselves for the consumption of others; to do this they must be *committed* to the deception, rather than just doing it by accident or as an expedient to a higher-level cooperative goal, as in the case of simplification in teaching. Thus we characterize cooperative domains as follows:

- They contain no committed deception,
- They contain no reports of intentional deception,
- They allow only first-hand reports of accidental deception (such reports are needed for repair dialogues).

Simulated agents for use in such domains can use a belief model in which only their own object level beliefs, their beliefs about others' object level beliefs (first-level nested beliefs) and their beliefs about others' beliefs about their own or others' object level beliefs (second-level nested beliefs) are recorded.

5 Mutual belief and infinite regress

We have seen that in order to take part in cooperative dialogue, a simulated agent need only represent beliefs at the object level and those at the first two levels of nesting. However, our illustration of the possibility of such a restriction requires only that the number of levels to be *distinguished* is three, and that the first three levels all be distinguished from each other. The illustration also points to deeper nesting levels of both belief and reasoning being in some way identical to the shallower ones in the cooperative case; if there were a PFPR it would correspond exactly to what happens when the plan for reaction is recognized. So rather than model agents that don't have deeply nested beliefs, we should design our belief model to accept new information, and respond to queries, about beliefs at *any* level of nesting, but allow deeper ones to be internally conflated with shallower ones.

In this section we shall look at the theoretical and practical advantages of such an approach, and describe a belief model in which all levels from doubly-nested downwards share a single representation of their nesting. Beliefs represented in this way will be referred to as *residual mutual beliefs (RMBs)* as when an agent has such a belief plus belief in the same proposition at the object level and that her conversant believes it, she has all the beliefs necessary to have mutual belief in it with her conversant. For agents S and A and proposition P, the proposition that S has RMB in P (SRMBP) is equivalent to the conjunction of SBABSBP, SBABSBABP, SBABSBABSBP, and so on. This characterization will be compared with the default state of Kobsa's BGP-MS belief model, i.e., its state when no extra differentiation between belief nestings has been made.

5.1 Reactive responses are a fiction

In section 4.3, we set out our view that the infinite regress entailed by Searle's definition of the process of understanding can be avoided by having the speaker make a plan which she intends to be recognized, but not actually plan at the propositional level the steps which constitute that recognition. In section 4.4.1 we added that indirect speech acts can be planned this way.

When this is done, the addressee is confronted with the problem of recognizing a plan that depends for its success on that very recognition. They can of course recognize it on the basis of the fiction that the speaker intended them to respond purely reactively, after which their actual response can differ if they decide that something other than a reactive response will in fact be more helpful to the speaker. But both the planning and plan recognition sides of the process look more consistent if we admit that what is actually happening is that the plan is being transferred bodily into the realms of mutual belief. This can happen because since the actual utterance is a co-presence event, so are the attitudes surrounding it. The speaker does not explicitly PFPR, nor does the addressee recognize such a move, but because the deeply-nested beliefs that these mental manœuvres would require are conflated with more shallowly nested beliefs, the maneuvers can be thought of as happening implicitly.

Putting this in terms of figure 2, what is happening is that the beliefs represented by the boxes in each line below the third are conflated with shallower beliefs, represented by boxes in the second and third lines, as shown in figure 3. Thus the reasoning processes represented by the lower lines are modelled implicitly when planning and plan recognition are carried out, but since they are unified with higher lines the plan *literally contains its own recognition*, so the need to suppose that at some level agents reason about fictitious reactive responses is removed.

5.2 Plans requiring mutual recognition

To further illustrate the reality and necessity of models with RMBs, let us consider a plan where the reactive stance simply doesn't work because there is nothing to react to. We shall argue that the generation and recognition of such plans requires the explicit manipulation of mutual beliefs, which can be represented if RMB is used as a component. Here is Fred walking towards the bus

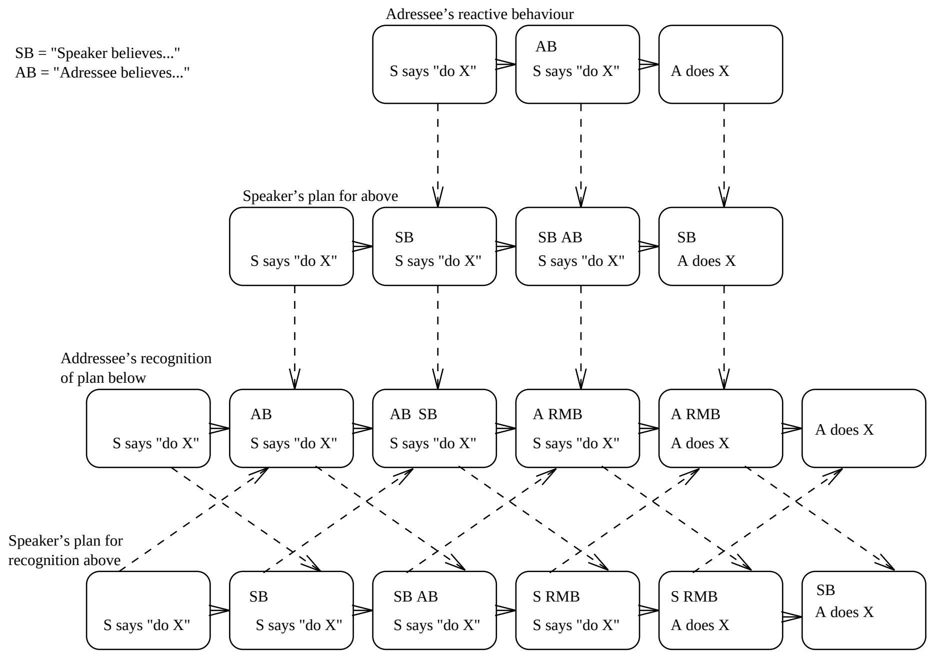

Figure 3: Self-modelling of reasoning in terms of RMBs

stop with the intention of catching the next bus. While still some way from the stop, he sees the bus approaching. He breaks into a run. Seeing this, Doris, the bus driver, stops at the stop and waits until Fred has arrived there and boarded before moving off. This could happen even if it was apparent to Doris that she would otherwise leave the bus stop before Fred arrived there, so merely running could not constitute part of a plan to catch the bus. Furthermore, running towards the stop isn't a conventional signal for communicating a message saying "Stop!", so from the point of view of both planning and plan recognition, running is an action which only achieves catching the bus if the driver waits.

One possibility is that Fred has a broken plan to board the bus, and Doris simply recognizes it and repairs it for him by waiting. This is unrealistic as Fred obviously wouldn't run if he didn't believe she would wait. So perhaps he is running because he thinks Doris will recognize his intention and wait? This too is problematic, because in Fred's model, Doris cannot see any reason for Fred to think she will wait, so concludes he is running for some other reason. If DW means "Doris waits", the problem is that DW required DBFBDW, which requires DBFBDBFBDW, and so forth.

We have to allow that at some deep level of nested reasoning, agents may mistakenly make or recognize plans. Suppose a very deeply nested belief such as DBFB . . . DBFBDW just popped into Doris' mind. Thanks to their mutual belief in her helpfulness, she could work out that she was expected to wait, and do so. However, imagining the process like this makes little sense; it merely illustrates that assuming DBMBDW is not far-fetched. If Doris makes that assumption, everything is explained, the shallower beliefs making up the MB being justified by the deeper ones.

The usefulness of assuming mutual belief directly also comes up in some examples of reference resolution. Imagine a version of the map task where the maps contain several similar trees as well as the regular landmarks. The IG, Doris, might say "Now go down to the tree" in a situation in which there were several trees the IF, Fred, could go down to, but one would make somewhat more sense in that context than the others. Although the reference does not identify it uniquely, the requirement of definite reference that *something* be identified uniquely causes the best available possible reference to be elevated to the status of a mutual belief. It works like the bus example: if T is the proposition that "The Tree" refers to the big one right below Fred, then FBT requires FBDBT, which in turn requires FBDBFBT (since Doris would not use the expression if she thought it might be misinterpreted) and so on. But since T is more likely than reference to any alternative tree, FBMBT is quite a plausible assumption to make, and, once made, is self-sustaining.

5.3 Predominance of situations affording mutual beliefs

Our final argument for the worth of specifying RMB in our belief model is tangential; the types of belief we can thus represent are precisely those that are afforded by the observations agents typically make. These observations may be private or carried out in the company of others, and in the latter case are referred to by Clark and Marshall (1981) as "co-presence events". Inferences can also be either private or co-present, the latter case depending on the fact that the information required to make them is the shared property of the culture from which the company making the inference are drawn. Direct observations and inferences can be treated the same way in what follows.

Agents can come by beliefs by doing any of the following:

- Observe or alter the domain and thus acquire object level beliefs (e.g., ABP)
- Observe other agents' interactions with the domain and thus acquire first-level nested beliefs in conjunction with object level beliefs (e.g., ABSBP).
- Take part in co-presence events and thus acquire mutual beliefs (ABP + ABSBP + ARMBP)

There are other conceivable combinations; agents could observe other agents observing them, or take part in co-present observations of individuals, and these would produce belief sets which would not fit neatly into our model. However, these and other more complex types of example have problems, i.e., when observing another one can't always see what *they* are observing, and in group observations of individuals it is usual for the individual to be in the group, and hence the whole thing to be a single level co-presence event.

These of course are not immutable limits; for any belief model short of a fully-expressive one there will be some types of observation that will afford beliefs of types with which it cannot cope. However, Clark and Marshall argue that the same applies to humans, and in section 6 we shall attempt to show that the model we have described will show degradation in just the sort of circumstances that could cause humans difficulty.

5.4 Grouping nested beliefs

We propose the use of a belief model in which the nesting levels of any possible belief identify it with one of three categories:

- The agent's private beliefs (SB)
- The agent's beliefs about her partner's beliefs (SBAB)
- The agent's beliefs about her partner's beliefs about what the agent believes to be mutually believed (SBABSBMB, or equivalently SRMB).

Models differentiating only these categories have the following desirable properties:

- They allow dialogue to be modelled in a way that avoids oversimplifying the process of understanding
- They support instances of comunication that depend upon mutual assumptions
- They provide a fair match to the combinations of beliefs that are afforded by observations and interactions in the world.

Note that this is not the only possible division; VIE-DPM (Kobsa 1985) referred to the following categories as being "primitives":

- The agent's private beliefs (SB)
- What the agent believes is mutually believed about her partner's beliefs (SBMBAB)
- What the agent believes her partner believes is mutually believed about her own beliefs (SBABMBSB)

Mutual beliefs about private beliefs he called "infinite-reflexive" beliefs. It is apparent that this grouping of beliefs would produce similar results in simulations to the one we discuss here, which has been chosen simply because it is easier to describe the examples in terms of a single category subsuming all deeply-nested beliefs. Other differences exist too: Kobsa's model is best suited to domains in which agents' knowledge is assumed to differ widely between conversants, such as user-modelling in database query systems. This is because the identities between beliefs that are conflated depend on ascription of nested beliefs rather than of object level beliefs. If the system believes a user believes P (SBUBP) it can ascribe its own belief about the user to the user, giving SBUBSBUBP, which is part of the same group. Similarly, if SBUBSBP, this too can be ascribed to the user to give SBUBSBUBSBP, again part of the same group as the original belief. Our model on the other hand is better suited to domains in which agents' knowledge is broadly similar, such as cooperative problem-solving. If we start with FBDBFBP, then Fred can model Doris modelling Fred ascribing his object level belief P to Doris to give FBDBFBDBP, and so on for the deeper levels of the RMB set.

Note that in either case we are assuming that agents' beliefs about their own beliefs are equivalent to the beliefs themselves, as is the case with epistemic logics consistent with S5 or KD45 (weak S5) modal logic.

5.5 Conclusion

In this section we have argued that in a belief model that groups beliefs into three sets according to level of nesting, beliefs at the third and deeper levels of nesting, rather than being ignored, should be identified with one of the shallower levels. The main argument for this is that it allows the model to capture accurately the concept of mutual belief, which is required to model recognition of a plan that involves that very recognition.

Other reasons for identifying deeper-nested beliefs with shallower ones are that situations agents are likely to come across tend to afford groups of such beliefs, and that modelling the process of belief ascription would in any case allow agents with one member of the set of nested beliefs that constitutes a group to derive the others.

Confusion between different beliefs within a group could potentially pose problems, but in section 6 we shall argue that these can usually be avoided, and that when they do crop up they are at least not an obstacle to communication with humans.

6 Beliefs and referring expressions

So far we have talked about modelling beliefs from the point of view of the relationship between an utterance and the plans of which it may form a part. These are beliefs related to the intentions of the speaker, and to the likelihood of the addressee adopting intentions as a result of the utterance. Interpreting an utterance, though, involves also the task of deciphering referring expressions, just as generating it involves choosing these in order that they can be correctly interpreted.

Referring expressions have been analysed in terms of the plan-based approach to dialogue (Appelt 1983) and have been used to argue for the reality of mutual beliefs amongst human conversants (Clark and Marshall 1981). In both arguments the possibility is brought up that individual deeply-nested beliefs play a role in their generation and interpretation. So far, our argument for limited-nesting models has rested on simplified, ambiguity-free cases. In this section we will attempt to extend the argument to more realistic cases.

In section 6.1 we argue that if a misunderstanding of a referring expression occurs between cooperative agents, they can recover from it without requiring a more complex belief model than that which we have described. This will be argued by reference to a more complex version of the map task. Section 6.2 considers problems that occur when we confront agents directly with information about deeply-nested beliefs about referent identities, and attempt to show that the sorts of mistakes and clarifications that would result from our belief model's inability to discriminate all such beliefs bears a resemblance to human behaviour in similar circumstances.

6.1 Cooperating to rectify ambiguity

Consider a version of the map task in which the agents can have multiple landmarks with the same name on their maps. This includes the possibility that even when the agents have agreed on a reference, they might each have identified it with a landmark in a different place. So when the IG makes the first mention of a landmark, the IF creates a new concept for it, expressing, e.g., "what my partner means by 'the swamp' ", abbreviated IG(SW0). If this can consistently designate a landmark for which the IF already knows a location, the reference is accepted and a belief IFB(IG(SW0) = SW1) added noting that they are indeed the same. There is still a need for both agents to be able to recover in the event that he has it wrong, so if the IF subsequently uses the same referring expression the IG, rather than simply treating it as a reference to the original concept, creates a concept for what the IF is referring to, IF(SW0), and adds a belief that it is indeed the original concept: IGB(IF(SW0) = SW0).

There are a number of ways in which referents that have been agreed on can later be found to be wrong, typically many at a time because of their interrelations. When the IF realizes she no longer has a consistent set of matches between individuals and referring expressions she has to indicate confusion and ask a question, e.g., "Sorry. Where is the swamp?" If this happens, the IG knows that the IF might have lost or changed other referents in the same adjustment. To give the best clarifying description, she therefore constructs a model of the IF's likely state of belief about all references. For this purposes she needs to represent concepts such as "What my partner thinks I mean by 'the swamp' ", abbreviated IF(IG(SW0)); these may be equated to similarly-named landmarks on her own map. Given that references are always chosen in order to be correctly understood, this is what the IG *will* be referring to if she now uses the expression "the swamp".

The IG's beliefs about such concepts are not explicitly nested beliefs, although it is apparent that such concepts are serving as notational shorthand for propositional second-level nested beliefs. For instance, if being fully acquainted with SW0 (having it "vivid" in JAM terms) implies knowing how far it is from the left of the map, then the IG having IF(IG(SW0)) vivid implies that she believes the IF believes she knows this distance. If we had to nest uncertainty about the identity of referents one more level, and talk about IG(IF(IG(SW0))) ("That which she thinks I think she is referring to"), we would be covertly introducing third-level nested beliefs.

Which levels we need depends on the way misunderstandings are recognized. If after the agents have used and accepted an expression, one, say the IG, realizes that they have been talking about different things, she must differentiate IF(SW1) from her own SW1. If she mentions the problem to the IF, he must differentiate IG(IF(SW0)) from IG(SW0). The IG must also differentiate IF(IG(SW1)) from IF(SW1), but does not need to distinguish IF(IG(IF(SW1))) from IF(SW1), because *before* the discovery of the misunderstanding the IF thought the IG had the same swamp as himself, and *afterwards* the IG knows the IF cannot have a guess at what the IG thinks he meant by "the swamp" that is more likely to be right than what he actually meant, since both are interpretations of the same information by the same agent. Indeed, because agents always use referring expressions they think will be understood correctly at the time, there is never any evidence to allow the third level to be differentiated from the first, and likewise the fourth from the second.

Note that in this case a belief model along the lines of Kobsa's primary partitions seems more appropriate than our own, as there are two separate groups of deep-nested beliefs – those at even and odd nesting levels. This is because the type of belief ascription assumed in Kobsa's model is more appropriate to referring expressions; if P is the proposition that "the swamp" denotes swamp SW0, then SBP implies SBABSBP, but not necessarily ABSBP. Conversely our model is more suited to beliefs about propositions in the domain; if P stands for the proposition that SW0 is at the bottom of the map, then SBP implies SBABP more directly than it does SBABSBP. The best model may be one that conflates beliefs of each kind in the appropriate way.

6.2 Dealing with deeply-nested assertions

When using a belief model that equates beliefs at different levels of nesting, we must be definite about what happens when the system is presented with evidence that should produce a belief at some but not all of these levels, in the case where the new beliefs would contradict what the system already believed at those levels. In our model the system should go into a state in which all deeply-nested beliefs about that proposition are considered uncertain. Obviously this, like any simplification, will result in some confusions; the best we can hope to say of our model is that these will be of a similar sort to those that humans suffer from when the situation affords deeply-nested beliefs with which they have difficulty coping.

To this end, let us consider Clark and Marshall's (1981) examples involving Ann and Bob discussing the movie showing that night. These examples describe a series of situations in which people cannot make certain definite references felicitously. They propose that in making a reference, a condition of full mutual knowledge is checked, and that this condition is specifically asserted by copresence events, such as discussing the movie over breakfast. Their examples illustrate the difficulty of referring to another movie when the conversants only believe at a finite series of levels of nesting that it is now the one to be shown. However it is also worth considering references to the original movie; that whose identity was originally the subject of mutual belief, but has been replaced at the object and shallowly-nested levels.

For instance, if Ann and Bob start out mutually knowing that the movie to be shown is "duck soup", following which Ann finds out that it has been changed, she will interpret a reference by Bob to "tonight's movie" as referring to "duck soup". However, in the example in which Ann knows that Bob knows the movie has been changed, but does not know he knows she knows it, we would expect Ann to be worried about referring to it and meaning the original one, despite the fact that if she knew Bob were perfectly rational and thought her to be the same, there would be no problem in doing so. In both cases Ann has RMB with Bob that the movie is "duck soup", but in the first case she also has first-level nested belief of this. Since in neither case can she mean the new movie, it appears that the beliefs checked to determine the usability of the referent are RMB plus possibly the corresponding first-level nested belief, but not the object level belief.

It may just be that conversants are doubtful about using referring expressions in this way because whatever happenstance leads to private beliefs differing from the RMB cannot be ruled out for the other conversant. For instance, if Ann knows Bob knows about the change in programme, how can she be sure that he is sure that she doesn't know about the change? Their cultural knowledge that the cinema will try to inform everyone of the change will affect all deeply-nested beliefs about the identity of the film, including those that make up RMB.

By contrast, if we picture a scenario in which each agent is convinced that the other has not come by any further information than they are known to have, RMB appears necessary and sufficient as a basis for plausible referring in cooperative situations. Jane could say to Ken: "The latest Ford Escort does 40 miles per gallon", even if she works at Ford and knows there is a top secret prototype of a newer Escort that does 50 miles per gallon. If P is the proposition that the newest Escort is the prototype, we have JBP, JBKB(not P) and JRMB(not P). Here Jane's confidence in the secrecy of the existence of the alternative referent allows her to use an "incorrect" one without anticipating problems. Now let us say that unbeknownst to Jane, Ken happens to be an industrial spy and has recently paid a clandestine visit to the Ford factory, on which he discovered the existence of the prototype and Jane's involvement with it. His relevant beliefs are thus KBP, KBJBP, KRMB(not P). Confident that he has not been detected, he can understand Jane's reference as referring to the currently available Escort, and could use the same reference the same way himself, even if he didn't mind Jane knowing he was a spy. But if Jane had watched Ken's visit on a video from a hidden security camera, she would at least have a moment's doubt when he mentioned the car, or before mentioning it herself. In this case, her beliefs in our model would be JBP and JBKBP; she would have no RMB either that P or (not P), as seeing Ken at work allows her to infer JBKBJBP, whereas JBKBJBKB(not P) and deeper-nested beliefs persist.

The result is that Jane cannot refer to either car simply as "the new Escort", and if Ken uses that expression she cannot tell which car he is talking about. In a cooperative dialogue situation, there are ways round this problem; using a more specific referring expression, or explicitly asking for clarification. This is at odds with the expected behaviour of a system with unlimited belief representation and unrestricted nested reasoning, which could repeatedly nest levels of PFPR until it was modelling a plan to refer in which the participants had full mutual knowledge of the referent's identity. Thus if there were any finite level of nesting beyond which all the beliefs were that (not P), such a system would use "the new Escort" to refer to the showroom model. Humans might carry out such reasoning to some degree in a non-cooperative situation. Jane and Ken might have goals of not giving away what they knew about one another, and would then use "the new Escort" to refer to the showroom model while watching for the slightest sign that the other was in any doubt about its meaning. Such behaviour would require at least some distinct representation of deeply nested private beliefs.

These examples give support to the idea that as long as there is no commitment to deception, people appear to check the conjunction of the series of beliefs that make up RMB (i.e., those at the second and deeper levels of nesting) for agreement on the likely interpretation of referring expressions. Clark & Marshall's examples show that finite nesting levels alone are not enough; the Escort examples add that all the deeper levels are not enough if the second level is missing, but the object and first levels are not critical. Note that the advantage the Kobsa model had in the map task ambiguity example has disappeared – where secrecy is assumed as in the Escort case, *no* belief ascriptions are sanctioned.

6.3 Conclusion

In domains in which there exists the potential for getting references wrong but agents have no prior knowledge about how other agents' choices of referent might differ from their own, a three-level model works as well as any, and should be sufficient for supporting the repair process. This is the case with ambiguities within the map task. In situations where agents do have some information to the effect that other agents' choices of referent for a given referring expression might differ from their own in specific ways (such as the movies and Escort examples), this information may be in the form of deeply-nested beliefs and hence not directly representable by our model. Thus our model will support the wrong behaviour, from the point of view of systems with fully-expressive belief models that reason unrestrictedly about one another on the basis that they share such models.

However, the behaviour produced by limited-nesting models (which essentially reason about one another as having one less level of nesting than themselves) may be more like human behaviour in cooperative situations. The suggestion is that humans, while able to handle deeply-nested beliefs to some extent, are "lazy" about doing so when cooperating, because of the opportunities to ask for clarification, or to choose their referring expressions to avoid the need for such reasoning. Hence a three-level model will be adequate for a system imitating human dialogue or even interacting with humans, where in fact unlimited models would produce unrealistic dialogue containing confusing referring expressions, and arouse human suspicions that deception was under way.

7 Consequences concerning dialogue simulation

We have seen that the belief states underlying dialogue in cooperative situations can be described in terms of the presence or absence of just three types of nested belief. Furthermore, the reasoning required for such dialogues can be restricted to object level inference, simple planning, and recognition of simple plans. This should enable simple, well-specified designs to be used for the reasoning and belief components of future cooperative AI dialogue systems.

7.1 Simulations usually assume cooperation

AI dialogue systems usually make the assumption of cooperation, either between their simulated conversants, or in the case of HCI systems, between the system and the human user. Since these systems make no attempt at being anything other than fully cooperative, they get by perfectly well with no deeply nested beliefs and no PFPR.

In fact, most existing systems combine one or more of the following characteristics:

- One conversant is subservient to the other in choosing goals,
- Nested beliefs are restricted to a certain maximum depth,
- The planning and/or reasoning components are incapable of indefinitely nested reasoning, i.e., an agent within the system models other agents' reasoning as being more restricted than its own.

The presence of any one of these restrictions reduces the amount of search that any system must do for language generation and understanding within a cooperative system. If we were to remove all these constraints from any dialogue system, the result would tend to generate and recognize a vast selection of plans for utterances, most of which will seem highly implausible to humans. For instance, when modelling a situation in which Doris is reasoning about Fred having asked her to open a door, the system will generate a number of possible beliefs for Doris, including:

- that Fred wants her to open the door,
- that some other agent has asked Fred to ask Doris to open the door,

- that Fred wants some other agent to know that he has asked Doris to open the door,
- that Fred wants her to go out of the room, and so forth.

A system with such capabilities must use a set of heuristics representing the ways in which language typically is and is not used if it is to choose the most likely of such interpretations. Each of the restrictions mentioned above, however, will prevent the more complex possible plans from being proposed.

7.2 Advantages of representing mutual belief explicitly

The most obvious reason for having an explicit representation for mutual belief in a dialogue system is that the direct establishment of mutual beliefs by co-presence events seems to provide the best model of what happens in human conversations. But it is not just a case of defining the deepest-nested beliefs in a system as embodying deeper-nested beliefs adding up to mutual beliefs; they must be imbued with the appropriate logical properties.

We argued in section 5.2 that such a belief model will allow some everyday nested reasoning situations to be modelled that otherwise would require unrealistic assumptions, e.g., allowing a plan to be recognized despite an obvious flaw. This ability requires the reasoning system to be able to assume RMB as a theory extension operator. Furthermore, a system whose deepest nested belief is considered to count as RMB will actually respond to new information justifying a conflicting belief at a deeper nesting level by ceasing to have RMB either way, rather than simply failing to capture the new information.

As for emulating the role of RMBs in allowing politeness-driven indirect utterances to be recognized using only straightforward plan recognition, this is something which requires further investigation, in particular into what sort of restricted belief model would be adequate to represent Brown and Levinson's constraining principles for polite requests.

7.3 Why not build a dialogue simulator that goes beyond cooperation?

The paradigm of representing utterances as having simple effects which are defeasible and can be reasoned about even if not actually occurring is sufficient for the generation and recognition of complex attempted deceptions, given unrestricted models of belief and reasoning. It might be assumed that deception is non-typical language use and therefore of little interest, but the same reasoning capabilities would also be needed to generate examples of tact (where this involves concealing beliefs) and morale-boosting (where this involves the promotion of false but beneficial beliefs). All these can be attempted within the restricted belief model we have described, but a more expressive belief model will result in improved strategies for such goals.

Apart from the question of whether we *want* to build HCI systems with such capabilities, providing them greatly increases the computational expense of generating and comprehending utterances, the space of possible plans growing faster than exponentially with the number of nesting levels being considered. It also increases the need for accurate descriptions of the constraints imposed on utterances by the requirements of politeness, correct implicature and relevance, so systems with unrestricted belief models may prove useful in investigating these phenomena.

8 Conclusion

Modelling conversants as rational agents has been a fruitful approach to the analysis of dialogue, and it appears that there are many more advances which can be made within this paradigm. In particular, it appears that we can produce models of complex dialogues such as the example from ViewGen, section 3.1.1, through the use of a representation of utterances as having simple but defeasible immediate effects, in combination with a reasoning system in which other agents are modelled as having the same logical capabilities as the agent modelling them. This approach is appealing because it treats communicative or linguistic action as a special case of rational action in general.

While it appears that such an agent model should be equipped with a belief database capable of representing nested beliefs down to any required level, such a belief model is too powerful in some respects and not powerful enough in others, and suffers serious problems of computational complexity. Individual deeply nested beliefs are not necessary beyond the second level of nesting, i.e., what an agent thinks another agent thinks a third agent (possibly the reasoning agent herself) thinks, as long as deception is not involved. However there are advantages in treating mutual belief as a primitive or a combination of a finite set of primitives in the representation.

Firstly, some important classes of dialogues which we want to model are characterized by the assumption of cooperative behaviour. This also applies to HCI systems, in which cooperation should be a requirement in the specification of the system. A dialogue is cooperative so long as neither participant has any salient goals which involve a commitment to withholding information from (i.e., not sharing their own beliefs with) any other agent, including the other conversant. With the assumption of cooperation, any goal that can be achieved by either agent through communication can in principle be achieved without nested reasoning beyond the second level. That is to say, an agent need only plan her behaviour on the basis that other agents will respond to it without reasoning about the motivations underlying it, and need only reason about other agents' behaviour with the presumption that it is planned in this way. Given that these restrictions apply to reasoning, there is no need for agents to represent nested beliefs beyond the second level.

Secondly, some kinds of communicative or cooperative action are most consistently explained by the existence of mutual beliefs other than as the conjunctions of individual deeply-nested beliefs. Such representations explain how actions that can only succeed when recognized and cooperated with can still succeed in the absence of any conventional communicative action requesting cooperation. They also allow the representation of utterances as having simple default effects to be used without the assumption at any point that agents actually act according to these effects.

Finally, we propose that belief representations explicitly include *residual mutual belief* i.e., mutual belief modulo object-level and singly-nested beliefs, the other two belief types which are themselves represented explicitly. Reasoning with RMBs can be thought of as second-level nested reasoning occurring simultaneously with deeper nested reasoning relating to the deeper individual nested beliefs subsumed within the RMB. Combining all deeply-nested beliefs into RMBs is especially appropriate for cooperative problem solving, as in such domains these are the beliefs that generally occur together due to belief ascription in any case.

References

Allen, J. F. (1983). Recognizing intentions from natural language utterances. In Brady and Berwick (Eds.), *Computational Models of Discourse*, pp. 107–166. Cambridge, MA: MIT Press.

Allen, J. F., L. K. Schubert, G. Ferguson, P. Heeman, C. H. Hwang, T. Kato, M. Light, N. G. Martin, B. W. Miller, M. Poesio, and D. R. Traum (1994). The TRAINS project: A case study in building a conversational agent. Technical Report 94-3, Computer Science Dept. University of Rochester, Rochester, New York.

Anderson, A. H., M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert (1991). The HCRC Map Task Corpus. *Language and Speech 34*(4), 351–366.

Appelt, D. E. (1983, October). Planning English referring expressions. Technical Report 312, Artificial Intelligence Center, SRI International.

Appelt, D. E. (1985). *Planning English Sentences.* Cambridge University Press.

Appelt, D. E. and K. Konolige (1988, April). A practical nonmonotonic theory for reasoning about speech acts. Technical Report 432, Artificial Intelligence Center, SRI International.

Appelt, D. E. and M. Pollack (1990). Weighted abduction as an inference method for plan recognition and evaluation. In *Second International Workshop on User Modeling*, Honolulu.

Ballim, A. and Y. Wilks (1991). *Artificial Believers.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Brown, P. and S. C. Levinson (1987). *Politeness: some universals in language usage* (second ed.). Cambridge University Press.

Carletta, J. C. (1992). *Risk-taking and Recovery in Task-Oriented Dialogue.* Ph. D. thesis, Edinburgh University Department of Artificial Intelligence.

Chapman, D. (1987). Planning for conjunctive goals. *Artificial Intelligence 32*(3), 333–377.

Clark, H. H. and C. R. Marshall (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, and I. A. Sag (Eds.), *Elements of Discourse Understanding*, pp. 10,63. Cambridge University Press.

Clark, H. H. and E. Schaefer (1989). Contributing to discourse. *Cognitive Science 13*, 259–294.

Cohen, P. and C. R. Perrault (1979). Elements of a plan-based theory of speech acts. *Cognitive Science 3*(3), 177–212.

Dunn, J. (1991). Understanding others: Evidence from naturalistic studies of children. In A. Whiten (Ed.), *Natural theories of mind: evolution, development and simulation of everyday mindreading.* Oxford: Basil Blackwell.

Fagin, R. and J. Y. Halpern (1988). Belief, awareness, and limited reasoning. *Artificial Intelligence 34*, 39–76.

Fikes, R. E. and N. J. Nilsson (1971). STRIPS: a new approach to the application of theorem proving to problem solving. *Artificial Intelligence 2*, 198–208.

Grice, H. P. (1957). Meaning. *Philosophical Review 66*, 377–88.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and Semantics 3: Speech Acts*, pp. 41–58. New York: Academic Press.

Halpern, J. Y. and Y. Moses (1985). A guide to the modal logics of knowledge and belief: Preliminary draft. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, Volume 1, pp. 480–490.

Hintikka, J. (1962). *Knowledge and Belief.* New York: Cornell University Press.

Houghton, G. (1986). *The Production of Language in Dialogue: A Computational Model.* Ph. D. thesis, University of Sussex.

Kautz, H. A. (1987). A formal theory of plan recognition. Technical Report TR-215, University of Rochester, Department of Computer Science, Rochester, New York.

Kobsa, A. (1985). VIE-DPM: A user model in a natural language dialogue system. In J. Laubsch (Ed.), *GWAI-84, 8th German Workshop on Artificial Intelligence*, Berlin. Springer.

Kobsa, A. (1990). Modeling the user's conceptual knowledge in BGP-MS, a user modelling shell system. *Computational Intelligence 6*, 193–208.

Konolige, K. (1986). *A deduction model of belief.* Los Altos, CA: Morgan Kaufman.

Leekam, S. R. (1991). Jokes and lies: Children's understanding of intentional falsehood. In A. Whiten (Ed.), *Natural theories of mind: evolution, development and simulation of everyday mindreading*. Oxford: Basil Blackwell.

Lewis, D. K. (1969). *Convention: a philosophical study*. Harvard University Press.

McDermott, D. V. (1982). A temporal logic for reasoning about processes and plans. *Cognitive Science 6*, 101–155.

Miller, G. A., E. Galanter, and K. H. Pribram (1960). *Plans and the structure of behavior*. New York: Holt.

Miller, P. H., F. S. Kessel, and J. H. Flavell (1970). Thinking about people thinking about people thinking about . . . : A study of social-cognitive development. *Child Development 41*, 613–623.

Perrault, C. R. (1990). An application of default logic to speech act theory. In P. R. Cohen, J. Morgan, and M. E. Pollack (Eds.), *Intentions in Communication*, pp. 161–185. Cambridge, MA: MIT Press.

Power, R. J. D. (1974). *A Computer Model of Conversation*. Ph. D. thesis, Edinburgh University Department of Artificial Intelligence.

Power, R. J. D. (1979). The organization of purposeful dialogue. *Linguistics 17*, 107–156.

Reichgelt, H. and N. Shadbolt (1989). Planning as theory extension. In *AISB89 Proceedings*, pp. 191–201. Pitman/Morgan Kaufmann.

Sacerdoti, E. D. (1975). The nonlinear nature of plans. In *Advance Papers of the Fourth International Joint Conference on Artificial Intelligence*, pp. 206–214.

Searle, J. R. (1969). *Speech acts: an essay in the philosophy of language*. Cambridge University Press.

Searle, J. R. (1975). Indirect speech acts. In P. Cole and J. L. Morgan (Eds.), *Syntax and Semantics 3: Speech Acts*, pp. 59–82. New York: Academic Press.

Taylor, J. (1993, October). An ATMS-based belief model for dialogue simulation. Technical Report HCRC/RP-45, Human Communication Research Centre, University of Edinburgh.

Taylor, J. (1994). *A Multi-Agent Planner for Modelling Dialogue*. Ph. D. thesis, School of Cognitive and Computing Sciences, University of Sussex.

Waldinger, R. (1977). Achieving several goals simultaneously. In E. W. Elcock and D. Michie (Eds.), *Machine Intelligence 8*. Halstead/Wiley.

Jasper Taylor received his PhD from the University of Sussex in 1994, and has been a researcher in the Human Communication Research Centre, University of Edinburgh, since 1992. Jean Carletta received her PhD from the University of Edinburgh in 1992 and has been a researcher in the Human Communication Research Centre, University of Edinburgh, since 1992. Chris Mellish received his PhD from the University of Edinburgh in 1981 and has been a lecturer in the Department of Artificial Intelligence, niversity of Edinburgh, since 1987.

Contents

List of Figures

List of Tables