A MODEL OF PLAN INFERENCE THAT DISTINGUISHES BETWEEN THE BELIEFS OF ACTORS AND OBSERVERS

Martha E. Pollack
Artificial Intelligence Center
and
Center for the Study of Language and Information
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025

ABSTRACT

Existing models of plan inference (PI) in conversation have assumed that the agent whose plan is being inferred (the actor) and the agent drawing the inference (the observer) have identical beliefs about actions in the domain. I argue that this assumption often results in failure of both the PI process and the communicative process that PI is meant to support. In particular, it precludes the principled generation of appropriate responses to queries that arise from invalid plans. I describe a model of PI that abandons this assumption. It rests on an analysis of plans as mental phenomena. Judgements that a plan is invalid are associated with particular discrepancies between the beliefs that the observer ascribes to the actor when the former believes that the latter has some plan, and the beliefs that the observer herself holds. I show that the content of an appropriate response to a query is affected by the types of any such discrepancies of belief judged to be present in the plan inferred to underlie that query. The PI model described here has been implemented in SPIRIT, a small demonstration system that answers questions about the domain of computer mail.

INTRODUCTION

The importance of plan inference (PI) in models of conversation has been widely noted in the computational-linguistics literature. Incorporating PI capabilities into systems that answer users' questions has enabled such systems to handle indirect speech acts [13], supply more information than is actually requested in a query [2], provide helpful information in response to a yes/no query answered in the negative [2], disambiguate requests [17], resolve certain forms of intersentential ellipsis [6,11], and handle such discourse phenomena as clarification subdialogues [11], and correction or "debugging" subdialogues [16,11].

The PI process in each of these systems, however, has assumed that the agent whose plan is being inferred (to whom I shall refer as the *actor*), and the agent drawing the inference (to whom I shall refer as the *observer*), have identical beliefs about the actions in the domain. Thus, Allen's model, which was one of the earliest accounts of PI in conversation[1] and inspired a great deal of the work done subsequently, includes, as a typical PI rule, the following: "$SBAW(P) \rightarrow_i SBAW(ACT)$ if P is a precondition of ACT" [2, page 120]. This rule can be glossed as "if the system (observer) believes that an agent (actor) wants some proposition P to be true, then the system may draw the inference that the agent wants to perform some action ACT of which P is a precondition." Note that it is left unstated precisely who it is—the observer or the actor—that believes that P is a precondition of ACT. If we take this to be a belief of the observer, it is not clear that the latter will infer the actor's plan; on the other hand, if we consider it to be a belief of the actor, it is unclear how the observer comes to have direct access to it. In practice, there is only a single set of operators relating preconditions and actions in Allen's system; the belief in question is regarded as being both the actor's and the observer's.

In many situations, an assumption that the relevant beliefs of the actor are identical with those of the observer results in failure not only of the PI process, but also of the communicative process that PI is meant to support. In particular, it precludes the principled generation of appropriate responses to queries that arise from invalid plans. In this paper, I report on a model of PI in conversation that distinguishes between the beliefs of the actor and those of the observer. The model rests on an analysis of plans as mental phenomena: "having a plan" is analyzed as having a particular configuration of beliefs and intentions. Judgements that a plan is invalid are associated with particular discrepancies between the beliefs that the observer ascribes to the actor when the former believes that the latter has some plan, and the beliefs observer herself holds. I give an account of different types of plan invalidities, and show how this account provides an explanation for certain regularities that are observable in cooperative responses to questions. The PI model described here has been implemented in SPIRIT, a small demonstration system that answers questions about the domain of computer mail. More

The research reported in this paper has been made possible in part by an IBM Graduate Fellowship, in part by a gift from the Systems Development Foundation, and in part by support from the Defense Advanced Research Projects Agency under Contract N00039-84-K-0078 with the Space and Naval Warfare Command. The views and conclusions contained in this document are those of the author and should not be interpreted as representative of the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the United States Government.

I am grateful to Barbara Grosz, James Allen, Phil Cohen, Amy Lansky, Candy Sidner and Bonnie Webber for their comments on an earlier draft.

[1] Allen's article [2] summarizes his dissertation research [1].

extensive discussion of both the PI model and SPIRIT can be found in my dissertation [14].

PLANS AS MENTAL PHENOMENA

We can distinguish between two views of plans. As Bratman [5, page 271] has observed, there is an ambiguity in speaking of an agent's plan: "On the one hand, [this] could mean an appropriate abstract structure—some sort of partial function from circumstances to actions, perhaps. On the other hand, [it] could mean an appropriate state of mind, one naturally describable in terms of such structures." We might call the former sense the *data-structure view of plans*, and the latter the *mental phenomenon view of plans*. Work in plan synthesis (e.g., Fikes and Nilsson [8], Sacerdoti [15], Wilkins [18], and Pednault [12]), has taken the data-structure view, considering plans to be structures encoding aggregates of actions that, when performed in circumstances satisfying some specified preconditions, achieve some specified results. For the purposes of PI, however, it is much more useful to adopt a mental phenomenon view and consider plans to be particular configurations of beliefs and intentions that some agent has. After all, inferring another agent's plan means figuring out what actions he "has in mind," and he may well be wrong about the effects of those intended actions.

Consider, for example, the plan I have to find out how Kathy is feeling. Believing that Kathy is at the hospital, I plan to do this by finding out the phone number of the hospital, calling there, asking to be connected to Kathy's room, and finally saying "How are you doing?" If, unbeknownst to me, Kathy has already been discharged, then executing my plan will not lead to my goal of finding out how she is feeling. For me to have a plan to do β that consists of doing some collection of actions Π, it is not necessary that the performance of Π actually lead to the performance of β. What is necessary is that I believe that its performance will do so. This insight is at the core of a view of plans as mental phenomena; in this view a plan "exists"—i.e., gains its status as a plan—by virtue of the beliefs, as well as the intentions, of the person whose plan it is.

Further consideration of our common-sense conceptions of what it means to have a plan leads to the following analysis [14, Chap. 3][2]:

(P0) An agent G has a plan to do β, that consists in doing some set of acts Π, provided that

1. G believes that he can execute each act in Π.
2. G believes that executing the acts in Π will entail the performance of β.
3. G believes that each act in Π plays a role in his plan. (See discussion below.)
4. G intends to execute each act in Π.
5. G intends to execute Π as a way of doing β.
6. G intends each act in Π to play a role in his plan.

The notion of an act *playing a role in* a plan is defined in terms of two relationships over acts: *generation*, in the sense defined by Goldman [9], and *enablement*. Roughly, one act *generates* another if, by performing the first, the agent also does the second; thus, saying to Kathy "How are you doing?" may generate asking her how she is feeling. Or, to take an example from the computer-mail domain, typing **DEL .** at the prompt for a computer mail system may generate deleting the current message, which may in turn generate cleaning out one's mail file. In contrast, one act *enables* the generation of a second by a third if the first brings about circumstances that are necessary for the generation. Thus, typing **HEADER 15** may enable the generation of deleting the fifteenth message by typing **DEL .**, because it makes message 15 be the current message, to which '.' refers.[3] The difference between generation and enablement consists largely in the fact that, when an act α generates an act β, the agent need only do α, and β will automatically be done also. However, when α enables the generation of some γ by β, the agent needs to do something more than just α to have done either β or γ. In this paper, I consider only the inference of a restricted subset of plans, which I shall call *simple plans*. An agent has a simple plan if and only if he believes that all the acts in that plan play a role in it by generating another act; i.e., if it includes no acts that he believes are related to one another by enablement.

It is important to distinguish between types of actions (act-types), such as typing **DEL .**, and actions themselves, such as my typing **DEL .** right now. Actions or acts—I will use the two terms interchangeably—can be thought of as triples of act-type, agent, and time. Generation is a relation over actions, not over act-types. Not every case of an agent typing **DEL .** will result in the agent deleting the current message; for example, my typing it just now did not, because I was not typing it to a computer mail system. Similarly, executability—the relation expressed in Clause (1) of (P0) as "can execute"—applies to actions, and the objects of an agent's intentions are, in this model, also actions.

Using the representation language specified in my thesis [14], which builds upon Allen's interval-based temporal logic [3], the conditions on G's having a simple plan to do β can be encoded as follows:

(P1) $\text{SIMPLE-PLAN}(G,\alpha_n,[\alpha_1,\ldots,\alpha_{n-1}],t_2,t_1) \leftrightarrow$

$$\begin{array}{ll}
\text{(i)} & \text{BEL}(G,\text{EXEC}(\alpha_i,G,t_2),t_1),\ \text{for } i = 1,\ldots,n \ \wedge \\
\text{(ii)} & \text{BEL}(G,\text{GEN}(\alpha_i,\alpha_{i+1},G,t_2),t_1),\ \text{for } i = 1,\ldots,n\text{-}1 \ \wedge \\
\text{(iii)} & \text{INT}(G,\alpha_i,t_2,t_1),\ \text{for } i = 1,\ldots,n \ \wedge \\
\text{(iv)} & \text{INT}(G,by(\alpha_i,\alpha_{i+1}),t_2,t_1),\ \text{for } i = 1,\ldots,n\text{-}1
\end{array}$$

The left-hand side of (P1) denotes that the agent G has, at time t_1, a simple plan to do α_n, consisting of doing the set of acts $\{\alpha_1,\ldots,\alpha_{n-1}\}$ at t_2. Note that all these are simultaneous acts; this is a consequence of the restriction to simple plans. The right-hand side of (P1) corresponds directly to (P0), except that, in keeping with the restriction to simple plans, specific assertions about each act generating another replace the

[2] Although this definition ignores some important issues of commitment over time, as discussed by Bratman [4] and Cohen and Levesque [7], it is sufficient to support the PI process needed for many question-answering situations. This is because, in such situations, unexpected changes in the world that would force a reconsideration of the actor's intentions can usually be safely ignored.

[3] Enablement here thus differs from the usual binary relation in which one action enables another. Since this paper does not further consider plans with enabling actions, the advantages of the alternative definition will not be discussed.

more general statement regarding the fact that each act plays a role in the plan. The relation BEL(G,P,t) should be taken to mean that agent G believes proposition P throughout time interval t; INT(G,α,t_2,t_1) means that at time t_1 G intends to do α at t_2. The relation EXEC(α,G,t) is true if and only if the act of G doing α at t is *executable*, and the relation GEN(α,β,G,t) is true if and only if the act of G doing α at t *generates* the act of G doing β at t. The function ***by*** maps two act-type terms into a third act-type term: if an agent G intends to do *by*(α,β), then G intends to do the complex act β-by-α, i.e., he intends to do α in order to do β. Further discussion of these relations and functions can be found in Pollack [14, Chap. 4].

Clause (i) of (P1) captures clause (1) of (P0).[4] Clause (ii) of (P1) captures both clauses (2) and (3) of (P0): when i takes the value n-1, clause (ii) of (P1) captures the requirement, stated in clause (2) of (P0), that G believes his acts will entail his goal; when i takes values between 1 and n-2, it captures the requirement of clause (3) of (P0), that G believes each of his acts plays a role in his plan. Similarly, clause (iii) of (P1) captures clause (4) of (P0), and clause (iv) of (P1) captures clauses (5) and (6) of (P0).

(P1) can be used to state what it means for an actor to have an invalid simple plan: G has an invalid simple plan if and only if he has the configuration of beliefs and intentions listed in (P1), where one or more of those beliefs is incorrect, and, consequently, one or more of the intentions is unrealizable. The correctness of the actor's beliefs thus determines the validity of his plan: if all the beliefs that are part of his plan are correct, then all the intentions in it are realizable, and the plan is valid. Validity in this absolute sense, however, is not of primary concern in modeling plan inference in conversation. What is important here is rather the observer's judgement of whether the actor's plan is valid. It is to the analysis of such invalidity judgements, and their effect on the question-answering process, that we now turn.

PLAN INFERENCE IN QUESTION-ANSWERING

Models of the question-answering process often include a claim that the respondent (R) must infer the plans of the questioner (Q). So R is the observer, and Q the actor. Building on the analysis of plans as mental phenomena, we can say that, if R believes that she has inferred Q's plan, there is some set of beliefs and intentions satisfying (P1) that R believes Q has (or is at least likely to have). Then there are particular discrepancies that may arise between the beliefs that R ascribes to Q when she believes he has some plan, and the beliefs that R herself holds. Specifically, R may not herself believe one or more of the beliefs, corresponding to Clauses (i) and (ii) of (P1), that she ascribes to Q. We can associate such discrepancies with R's judgement that the plan she has inferred is invalid.[5] The type of any invalidities, defined in terms of the clauses of (P1) that contain the discrepant beliefs, can be shown to influence the content of a cooperative response. However, they do not fully determine it: the plan inferred to underlie a query, along with any invalidities it is judged to have, are but two factors affecting the response-generation process, the most significant others being factors of relevance and salience.

I will illustrate the effect of invalidity judgements on response content with a query of the form "I want to perform an act of β, so I need to find out how to perform an act of α," in which the goal is explicit, as in example (1) below[6]:

(1) "I want to prevent Tom from reading my mail file. How can I set the permissions on it to faculty-read only?"

In questions in which no goal is mentioned explicitly, analysis depends upon inferring a plan leading to a goal that is reasonable in the domain situation. Let us assume that, given query (1), R has inferred that Q has the simple plan that consists only in setting the permissions to faculty-read only, and thereby directly preventing Tom from reading the file, i.e.:

(2)BEL(R,SIMPLE-PLAN(Q, prevent(mmfile,read,tom),
[set-permissions(mmfile,read,faculty)],
t_2, t_1),
t_1)

Later in this paper, I will describe the process by which R can come to have this belief. Bear in mind that, by (P1), (2) can be expanded into a set of beliefs that R has about Q's beliefs and intentions.

The first potential discrepancy is that R may believe to be false some belief, corresponding to Clause (i) of (P1), that, by virtue of (2), she ascribes to Q. In such a case, I will say that she believes that some action in the inferred plan is *unexecutable*. Examples of responses in which R conveys this information are (3) (in which R believes that at least one intended act is unexecutable) and (4) (in which R believes that at least two intended acts are unexecutable):

(3) "There is no way for you to set the permissions on a file to faculty-read only. What you can do is move it into a password-protected subdirectory; that will prevent Tom from reading it."

(4) "There is no way for you to set the permissions on a file to faculty-read only, nor is there any way for you to prevent Tom from reading it."

[4] In fact, it captures more: to encode Clause (i) of (P0), the parameter i in Clause (i) of (P1) need only vary between 1 and n-1. However, given the relationship between EXEC and GEN specified in Pollack [14], namely

$$EXEC(\alpha,G,t) \land GEN(\alpha,\beta,G,t) \rightarrow EXEC(\beta,G,t)$$

the instance of Clause (i) of (P1) with i=n is a consequence of the instance of Clause (i) with i=n-1 and the instance of Clause (ii) with i=n-1. A similar argument can be made about Clause (iii).

[5] This assumes that R always believes that her own beliefs are complete and correct. Such an assumption is not an unreasonable one for question-answering systems to make. More general conversational systems must abandon this assumption, sometimes updating their own beliefs upon detecting a discrepancy.

[6] The analysis below is related to that provided by Joshi, Webber, and Weischedel [10]. There are significant differences in my approach, however, which involve (i) a different structural analysis, which applies *unexecutability* to actions rather than plans and introduces *incoherence* (this latter notion I define in the next section); (ii) a claim that the types of invalidities (e.g., formedness, executability of the queried action, and executability of a goal action) are independent of one another; and (iii) a claim that recognition of any invalidities, while necessary for determining what information to include in an appropriate response, is not in itself sufficient for this purpose. Also, Joshi et al. do not consider the question of how invalid plans can be inferred.

The discrepancy resulting in (3) is represented in (5); the discrepancy in (4) is represented in (5) plus (6):

$$\begin{aligned}
&(5)\text{BEL}(\text{R},\text{BEL}(\text{Q},\text{EXEC}(\text{set-permissions}(\text{mmfile},\text{read},\text{faculty}),\\
&\qquad\qquad\qquad\qquad\quad \text{Q},t_2),\\
&\qquad\qquad t_1),\\
&\qquad t_1)\\
&\qquad \wedge\\
&\quad \text{BEL}(\text{R},\neg\text{EXEC}(\text{set-permissions}(\text{mmfile},\text{read},\text{faculty}),\\
&\qquad\qquad\qquad\quad \text{Q},t_2),\\
&\qquad t_1)
\end{aligned}$$

$$\begin{aligned}
&(6)\text{BEL}(\text{R},\text{BEL}(\text{Q},\text{EXEC}(\text{prevent}(\text{mmfile},\text{read},\text{tom}),\\
&\qquad\qquad\qquad\qquad\quad \text{Q},t_2),\\
&\qquad\qquad t_1),\\
&\qquad t_1)\\
&\qquad \wedge\\
&\quad \text{BEL}(\text{R},\neg\text{EXEC}(\text{prevent}(\text{mmfile},\text{read},\text{tom}),\\
&\qquad\qquad\qquad\quad \text{Q},t_2),\\
&\qquad t_1)
\end{aligned}$$

The second potential discrepancy is that R may believe false some belief corresponding to Clause (ii) of (P1) that, by virtue of (2), she ascribes to Q. I will then say that she believes the plan to be *ill-formed.* In this case, her response may convey that the intended acts in the plan will not fit together as expected, as in (7), which might be uttered if R believes it to be mutually believed by R and Q that Tom is the system manager:

(7) "Well, the command is **SET PROTECTION = (Faculty:Read)**, *but that won't keep Tom out: file permissions don't apply to the system manager."*

The discrepancy resulting in (7) is (8):

$$\begin{aligned}
&(8)\text{BEL}(\text{R},\text{BEL}(\text{Q},\text{GEN}(\text{set-permissions}(\text{mmfile},\text{read},\text{faculty}),\\
&\qquad\qquad\qquad\qquad\quad \text{prevent}(\text{mmfile},\text{read},\text{tom}),\\
&\qquad\qquad\qquad\qquad\quad \text{Q},t_2),\\
&\qquad\qquad t_1),\\
&\qquad t_1)\\
&\qquad \wedge\\
&\quad \text{BEL}(\text{R},\neg\text{GEN}(\text{set-permissions}(\text{mmfile},\text{read},\text{faculty}),\\
&\qquad\qquad\qquad\quad \text{prevent}(\text{mmfile},\text{read},\text{tom}),\\
&\qquad\qquad\qquad\quad \text{Q},t_2),\\
&\qquad t_1)
\end{aligned}$$

Alternatively, there may be some combination of these discrepancies between R's own beliefs and those that R attributes to Q, as reflected in a response such as (9):

(9) "There is no way for you to set the permissions to faculty-read only; and even if you could, it wouldn't keep Tom out: file permissions don't apply to the system manager."

The discrepancies encoded in (5) and (8) together might result in (9).

Of course, it is also possible that no discrepancy exists at all, in which case I will say that R believes that Q's plan is *valid.* A response such as (10) can be modeled as arising from an inferred plan that R believes valid:

(10) "Type **SET PROTECTION = (Faculty:Read)**.*"*

Of the eight possible combinations of formedness, executability of the queried act and executability of the goal act, seven are possible: the only logically incompatible combination is a well-formed plan with an executable queried act, but unexecutable goal act. This range of invalidities accounts for a great deal of the information conveyed in naturally occurring dialogues. But there is an important regularity that the PI model does not yet explain.

A PROBLEM FOR PLAN INFERENCE

In all of the preceding cases, R has intuitively "made sense" of Q's query, by determining some underlying plan whose components she understands, though she may also believe that the plan is flawed. For instance in (7), R has determined that Q may mistakenly believe that, when one sets the permissions on a file to allow a particular access to a particular group, no one who is not a member of that group can gain access to the file. This (incorrect) belief explains why Q believes that setting the permissions will prevent Tom from reading the file.

There are also cases in which R may not even be able to "make sense" of Q's query. As a somewhat whimsical example, imagine Q saying:

(11) "I want to talk to Kathy, so I need to find out how to stand on my head."

In many contexts, a perfectly reasonable response to this query is "Huh?". Q's query is *incoherent*: R cannot understand why Q believes that finding out how to stand on his head (or standing on his head) will lead to talking with Kathy. One can, of course, construct scenarios in which Q's query makes perfect sense: Kathy might, for example, be currently hanging by her feet in gravity boots. The point here is not to imagine such circumstances in which Q's query would be coherent, but instead to realize that there are many circumstances in which it would not.

The judgement that a query is incoherent is not the same as a judgement that the plan inferred to underlie it is ill-formed. To see this, contrast example (11) with the following:

(12) "I want to talk to Kathy. Do you know the phone number at the hospital?"

Here, if R believes that Kathy has already been discharged from the hospital, she may judge the plan she infers to underlie Q's query to be ill-formed, and may inform him that calling the hospital will not lead to talking to Kathy. She can even inform him why the plan is ill-formed, namely, because Kathy is no longer at the hospital. This differs from (11), in which R cannot inform Q of the reason his plan is invalid, because she cannot, on an intuitive level, even determine what his plan is.

Unfortunately, the model as developed so far does not distinguish between incoherence and ill-formedness. The reason is that, given a reasonable account of semantic interpretation, it is transparent from the query in (11) that Q intends to talk to Kathy, intends to find out how to stand on his head, and intends his doing the latter to play a role in his plan to do the former and that he also believes that he can talk to Kathy, believes that he can find out how to stand on his head, and believes that his doing the latter will play a role in his plan to do the former.[7] But these beliefs and intentions are precisely what are required to have a plan according to (P0). Consequently, after hearing (11), R can, in fact, infer a plan underlying Q's query, namely the obvious one: to find out how to stand on his head (or to stand on his head) in order to talk to Kathy. Then, since R does not herself believe that the former act will lead to the latter, on the analysis so far given, we would regard R as judging Q's plan to be ill-formed. But this is not the desired analysis: the model should instead capture the fact that R cannot make sense of Q's query here—that it is incoherent.

Let us return to the set of examples about setting the permissions on a file, discussed in the previous section. In her semantic interpretation of the query in (1), R may come to have a number of beliefs about Q's beliefs and intentions. Specifically, all of the following may be true:

$$\begin{array}{l}(13)\mathrm{BEL}(\mathrm{R},\mathrm{BEL}(\mathrm{Q},\mathrm{EXEC}(\text{set-permissions}(\text{mmfile},\text{read},\text{faculty}),\\ \qquad\qquad\qquad\qquad \mathrm{Q},t_2),\\ \qquad\qquad t_1),\\ \qquad t_1)\end{array}$$

$$\begin{array}{l}(14)\mathrm{BEL}(\mathrm{R},\mathrm{BEL}(\mathrm{Q},\mathrm{EXEC}(\text{prevent}(\text{mmfile},\text{read},\text{tom}),\\ \qquad\qquad\qquad\qquad \mathrm{Q},t_2),\\ \qquad\qquad t_1),\\ \qquad t_1)\end{array}$$

$$\begin{array}{l}(15)\mathrm{BEL}(\mathrm{R},\mathrm{BEL}(\mathrm{Q},\mathrm{GEN}(\text{set-permissions}(\text{mmfile},\text{read},\text{faculty}),\\ \qquad\qquad\qquad\qquad \text{prevent}(\text{mmfile},\text{read},\text{tom}),\\ \qquad\qquad\qquad\qquad \mathrm{Q},t_2),\\ \qquad\qquad t_1),\\ \qquad t_1)\end{array}$$

$$\begin{array}{l}(16)\mathrm{BEL}(\mathrm{R},\mathrm{INT}(\mathrm{Q},\text{set-permissions}(\text{mmfile},\text{read},\text{faculty}),\\ \qquad\qquad t_2,t_1),\\ \qquad t_1)\end{array}$$

$$\begin{array}{l}(17)\mathrm{BEL}(\mathrm{R},\mathrm{INT}(\mathrm{Q},\text{prevent}(\text{mmfile},\text{read},\text{tom}),\\ \qquad\qquad t_2,t_1),\\ \qquad t_1)\end{array}$$

$$\begin{array}{l}(18)\mathrm{BEL}(\mathrm{R},\mathrm{INT}(\mathrm{Q},\mathit{by}(\text{set-permissions}(\text{mmfile},\text{read},\text{faculty}),\\ \qquad\qquad\qquad \text{prevent}(\text{mmfile},\text{read},\text{tom})),\\ \qquad\qquad t_2,t_1),\\ \qquad t_1)\end{array}$$

Together, (13)-(18) are sufficient for R's believing that Q has the simple plan as expressed in (2). This much is not surprising. In effect Q has stated in his query what his plan is—to prevent Tom from reading the file by setting the permission on it to faculty-read only—so, of course, R should be able to infer just that. And if R further believes that the system manager can override file permissions and that Tom is the system manager, but also that Q does not know the former fact, R will judge that Q's plan is ill-formed, and may provide a response such as that in (7). There is a discrepancy here between the belief R ascribes to Q in satisfaction of Clause (ii) of (P1)—namely, that expressed in (15)—and R's own beliefs about the domain.

But what if R, instead of believing that it is mutually believed by Q and R that Tom is the system manager, believes that they mutually believe that he is a faculty member? In this case, (13)-(18) may still be true. However we do not want to say that this case is indistinguishable from the previous one. In the previous case, R understood the source of Q's erroneous belief: she realized that Q did not know that the system manager could override file protections, and therefore thought that, by setting permissions to restrict access to a group that Tom is not a member of, he could prevent Tom from reading the file. In contrast, in the current case, R cannot really understand Q's plan: she cannot determine why Q believes that he will prevent Tom from reading the file by setting the permissions on it to faculty-read only, given that Q believes that Tom is a faculty member. This current case is like the case in (11): Q's query is incoherent to R.

To capture the difference between ill-formedness and incoherence, I will claim that, when an agent R is asked a question by an actor Q, R needs to attempt to ascribe to Q more than just a set of beliefs and intentions satisfying (P1). Specifically, for each belief satisfying Clause (ii) of (P1), R must also ascribe to Q another belief that explains the former in a certain specifiable way. The beliefs that satisfy Clause (ii) are beliefs about the relation between two particular actions: for instance, the plan underlying query (12) includes Q's belief that his action of calling the hospital at t_2 will generate his action of establishing a communication channel to Kathy at t_2. This belief can be explained by a belief Q has about the relation between the act-types "calling a location" and "establishing a communication channel to an agent." Q may believe that acts of the former type generate acts of the latter type provided that the agent to whom the communication channel is to be established is at the location to be called. Such a belief can be encoded using the predicate CGEN, which can be read "conditionally generates," as follows:

$$(19)\mathrm{BEL}(\mathrm{Q},\ \mathrm{CGEN}(\text{call}(\mathrm{X}),\text{establish-channel}(\mathrm{Y}),\text{at}(\mathrm{X},\mathrm{Y})),\ t_1)$$

The relation $\mathrm{CGEN}(\alpha,\beta,C)$ is true if and only if acts of type α performed when condition C holds will generate acts of type β. Thus, the sentence $\mathrm{CGEN}(\alpha,\beta,C)$ can be seen as one possible interpretation of a hierarchical planning operator with header β, preconditions C, and body α. Conditional generation is a relation between two act-types and a set of conditions; generation, which is a relation between two actions, can be defined in terms of conditional generation.

In reasoning about (12), R can attribute to Q the belief expressed in (19), combined with a belief that Kathy will be at the hospital at time t_2. Together, these beliefs explain Q's belief that, by calling the hospital at t_2, he will establish a communication channel to Kathy. Similarly, in reasoning about query (1) in the case in which R does not believe that Q knows that Tom is a faculty member, R can ascribe to Q the beliefs that, by setting the permissions on a file to restrict access to a particular group, one denies access to everyone who is neither a member of that group nor the system manager, as expressed in (20):

$$\begin{array}{l}(20)\mathrm{BEL}(\mathrm{R},\mathrm{BEL}(\mathrm{Q},\mathrm{CGEN}(\text{set-permissions}(\mathrm{X},\mathrm{P},\mathrm{Y}),\\ \qquad\qquad\qquad\qquad \text{prevent}(\mathrm{X},\mathrm{P},\mathrm{Z}),\\ \qquad\qquad\qquad\qquad \neg\text{member}(\mathrm{Z},\mathrm{Y})),\\ \qquad\qquad t_1),\\ \qquad t_1)\end{array}$$

She can also ascribe to Q the belief that Tom is not a member of the faculty, (or more precisely, that Tom will not be a member of the faculty at the intended performance time t_2), i.e.,

[7] Actually, the requirement that Q have these beliefs may be slightly too strong; see Pollack [14, Chap. 3] for discussion.

(21)$\mathrm{BEL}(R,\mathrm{BEL}(Q,\ \mathrm{HOLDS}(\neg\mathrm{member}(tom,faculty),t_2),t_1),t_1)$

The conjunction of these two beliefs explains Q's further belief, expressed in (15), that, by setting the permissions to faculty-read only at t_2, he can prevent Tom from reading the file.

In contrast, in example (11), R has no basis for ascribing to Q beliefs that will explain why he thinks that standing on his head will lead to talking with Kathy. And, in the version of example (1) in which R believes that Q believes that Tom is a faculty member, R has no basis for ascribing to Q a belief that explains Q's belief that setting the permissions to faculty-read only will prevent Tom from reading the file.

Explanatory beliefs are incorporated in the PI model by the introduction of *explanatory plans*, or *eplans*. Saying that an agent R believes that another agent Q has some eplan is shorthand for describing a set of beliefs possessed by R, specifically:

(P2) $(R,\mathrm{EPLAN}(Q,\alpha_n,[\alpha_1,\ldots,\alpha_{n-1}],[\rho_1,\ldots,\rho_{n-1}],$
$t_2,t_1),t_1)$
$\leftrightarrow$

(i) $\mathrm{BEL}(R,\mathrm{BEL}(Q,\mathrm{EXEC}(\alpha_i,Q,t_2),t_1),t_1)$,
for $i = 1,\ldots,n \wedge$

(ii) $\mathrm{BEL}(R,\mathrm{BEL}(Q,\mathrm{GEN}(\alpha_i,\alpha_{i+1},Q,t_2),t_1),t_1)$,
for $i = 1,\ldots,n\text{-}1 \wedge$

(iii) $\mathrm{BEL}(R,\mathrm{INT}(Q,\alpha_i,t_2,t_1),t_1)$,
for $i = 1,\ldots,n \wedge$

(iv) $\mathrm{BEL}(R,\mathrm{INT}(Q,by(\alpha_i,\alpha_{i+1}),\ t_2,t_1),t_1)$,
for $i = 1,\ldots,n\text{-}1 \wedge$

(v) $\mathrm{BEL}(R,\mathrm{BEL}(Q,\rho_i,t_1),t_1)$,
where each ρ_i is
$\mathrm{CGEN}(\alpha_i,\alpha_{i+1},C_i) \wedge \mathrm{HOLDS}(C_i,t_2)$

I claim that the PI process underlying cooperative question-answering can be modeled as an attempt to infer an eplan, i.e., to form a set of beliefs about the questioner's beliefs and intentions that satisfies (P2). Thus the next question to ask is: how can R come to have such a set of beliefs?

THE INFERENCE PROCESS

In the complete PI model, the inference of an eplan is a two-stage process. First, R infers beliefs and intentions that Q plausibly has. Then when she has found some set of these that is large enough to account for Q's query, their epistemic status can be upgraded, from beliefs and intentions that R believes Q plausibly has, to beliefs and intentions that R will, for the purposes of forming her response, consider Q actually to have. Within this paper, however, I will blur the distinction between attitudes that R believes Q plausibly has and attitudes that R believes Q indeed has; in consequence I will also omit discussion of the second stage of the PI process.

A set of plan inference rules encodes the principles by which an inferring agent R can reason from some set of beliefs and intentions—call this the antecedent eplan—that she thinks Q has, to some further set of beliefs and intentions—call this the consequent eplan—that she also thinks he has. The beliefs and intentions that the antecedent eplan comprises are a proper subset of those that the consequent eplan comprises. To reason from antecedent eplan to consequent eplan, R must attribute some explanatory belief to Q on the basis of something other than just Q's query. In more detail, if part of R's belief that Q has the antecedent eplan is a belief that Q intends to do some act α, and R has reason to believe that Q believes that act-type α conditionally generates act-type γ under condition C, then R can infer that Q intends to do α in order to do γ, believing as well that C will hold at performance time. R can also reason in the other direction: if part of her belief that Q has some plausible eplan is a belief that Q intends to do some act α and R has reason to believe that Q believes that act-type γ conditionally generates act-type α under condition C, then R can infer that Q intends to do γ in order to do α, believing that C will hold at performance time.

The plan inference rules encode the pattern of reasoning expressed in the last two sentences. Different plan inference rules encode the different bases upon which R may decide that Q may believe that a conditional generation relation holds between some α, an act of which is intended as part of the antecedent eplan, and some γ. This ascription of beliefs, as well as the ascription of intentions, is a nonmonotonic process. For arbitrary proposition P, R will only decide that Q may believe that P if R has no reason to believe Q believes that $\neg P$.

In the most straightforward case, R will ascribe to Q a belief about a conditional generation relation that she herself believes true. This reasoning can be encoded in the representation language in rule (PI1):

(PI1) $\mathrm{BEL}(R,\mathrm{EPLAN}(Q,\alpha_n,[\alpha_1,\ldots,\alpha_{n-1}],[\rho_1,\ldots,\rho_{n-1}],$
$t_2,t_1),t_1)$
$\wedge$
$\mathrm{BEL}(R,\mathrm{CGEN}(\alpha_n,\gamma,C),t_1)$
$\rightarrow$
$\mathrm{BEL}(R,\mathrm{EPLAN}(Q,\gamma,[\alpha_1,\ldots,\alpha_n],[\rho_1,\ldots,\rho_n],t_2,t_1),t_1)$
where $\rho_n = CGEN(\alpha_n,\gamma,C) \wedge HOLDS(C,t_2)$

This rule says that, if R's belief that Q has some eplan includes a belief that Q intends to do an act α_n, and R also believes that act-type α_n conditionally generates some γ under condition C, then R can (nonmonotonically) infer that Q has the additional intention of doing α_n in order to do γ—i.e., that he intends to do $by(\alpha_n,\gamma)$. Q's having this intention depends upon his also having the supporting belief that α_n conditionally generates γ under some condition C, and the further belief that this C will hold at performance time. A rule symmetric to (PI1) is also needed since R can not only reason about what acts might be generated by an act that she already believes Q intends, but also about what acts might generate such an act.

Consider R's use of (PI1) in attempting to infer the plan underlying query (1).[8] R herself has a particular belief about the relation between the act-types "setting the permissions on a file" and "preventing someone access to the file," a belief we can encode as follows:

(22)$\mathrm{BEL}(R,\mathrm{CGEN}(\text{set-permissions}(X,P,Y),$
$\text{prevent}(X,P,Z),$
$\neg\text{member}(Z,Y) \wedge \neg\text{system-mgr}(Z)),$
$t_1)$

From query (1), R can directly attribute to Q two trivial eplans:

[8] I have simplified somewhat in the following account for presentational purposes. A step-by-step account of this inference process is given in Pollack [14, Chap. 6].

(23)BEL(R,EPLAN(Q,set-permissions(mmfile,read,facult y),
[ ],t_2, t_1),
t_1)
(24)BEL(R,EPLAN(Q,prevent(mmfile,read,tom),[ ],t_2, t_1),
t_1)

The belief in (23) is justified by the fact that (13) satisfies Clause (i) of (P2), (16) satisfies Clause (iv) of (P2), and Clauses (ii), (iii), and (v) are vacuously satisfied. An analogous argument applies to (24).

Now, if R applies (PI1), she will attribute to Q exactly the same belief as she herself has, as expressed in (22), along with a belief that the condition C specified there will hold at t_2. That is, as part of her belief that a particular eplan underlies (1), R will have the following belief:

(25)BEL(R,BEL(Q,CGEN(set-permissions(X,P,Y),
prevent(X,P,Z),
$\neg$member(Z,Y) $\wedge$ $\neg$system-mrg(Z))
$\wedge$
HOLDS($\neg$member(tom,faculty)
$\wedge$ $\neg$system-mgr(tom), t_2),
t_1),
t_1)

The belief that R attributes to Q, as expressed in (25), is an explanatory belief supporting (15). Note that it is not the same explanatory belief that was expressed in (20) and (21). In (25), the discrepancy between R's beliefs and R's beliefs about Q's beliefs is about whether Tom is the system manager. This discrepancy may result in a response like (26), which conveys different information than does (7) about the source of the judged ill-formedness.

(26) "Well, the command is **SET PROTECTION = (Faculty:Read)**, *but that won't keep Tom out: he's the system manager."*

(PI1) (and its symmetric partner) are not sufficient to model the inference of the eplan that results in (7). This is because, in using (PI1), R is restricted to ascribing to Q the same beliefs about the relation between domain act-types as she herself has.[9] The eplan that results in (7) includes a belief that R attributes to Q involving a relation between act-types that R believes false, specifically, the CGEN relation in (20). What is needed to derive this is a rule such as (PI2):

(PI2) BEL(R,EPLAN(Q,α_n,[$\alpha_1, \ldots, \alpha_{n-1}$],[$\rho_1, \ldots, \rho_{n-1}$],
t_2, t_1),t_1)
$\wedge$
BEL(R,CGEN($\alpha_n, \gamma, C_1 \wedge \ldots \wedge C_m$),$t_1$)
$\longrightarrow$
BEL(R,EPLAN(Q,γ,[$\alpha_1, \ldots, \alpha_n$],[$\rho_1, \ldots, \rho_n$],$t_2, t_1$),$t_1$)
where $\rho_n = CGEN(\alpha_n, \gamma, C_1 \wedge \ldots \wedge C_{i-1} \wedge C_{i+1} \wedge \ldots \wedge C_m) \wedge$
$HOLDS(C_1 \wedge \ldots \wedge C_{i-1} \wedge C_{i+1} \wedge \ldots \wedge C_m, t_2)$

What (PI2) expresses is that R may ascribe to Q a belief about a relation between act-types that is a slight variation of one she herself has. What (PI2) asserts is that, if there is some CGEN relation that R believes true, she may attribute to Q a belief in a similar CGEN relation that is stronger, in that it is missing one of the required conditions. If R uses (PI2) in attempting to infer the plan that underlies query (1), she may decide that Q's belief about the conditions under which setting the permissions on a file prevents someone from accessing the file do not include the person's not being the system manager. This can result in R attributing to Q the explanatory belief in (20) and (21), which, in turn, may result in a response such as that in (7).

Of course, both the kind of discrepancy that may be introduced by (PI1) and the kind that is always introduced by (PI2) may be present simultaneously, resulting in a response like (27):

(27) "Well, the command is **SET PROTECTION = (Faculty:Read)**, *but that won't keep Tom out: he's the system manager, and file permissions don't apply to the system manager."*

(PI2) represents just one kind of variation of her own beliefs that R may consider attributing to Q. Additional PI rules encode other variations and can also be used to encode any typical misconceptions that R may attribute to Q.

IMPLEMENTATION

The inference process described in this paper has been implemented in SPIRIT, a System for Plan Inference that Reasons about Invalidities Too. SPIRIT infers and evaluates the plans underlying questions asked by users about the domain of computer mail. It also uses the result of its inference and evaluation to generate simulated cooperative responses. SPIRIT is implemented in C-Prolog, and has run on several different machines, including a Sun Workstation, a Vax 11-750, and a DEC-20. SPIRIT is a demonstration system, implemented to demonstrate the PI model developed in this work; consequently only a few key examples, which are sufficient to demonstrate SPIRIT's capabilities, have been implemented. Of course, SPIRIT's knowledge base could be expanded in a straightforward manner. SPIRIT has no mechanisms for computing relevance or salience and, consequently, always produces as complete an answer as possible.

CONCLUSION

In this paper I demonstrated that modeling cooperative conversation, in particular cooperative question-answering, requires a model of plan inference that distinguishes between the beliefs of actors and those of observers. I reported on such a model, which rests on an analysis of plans as mental phenomena. Under this analysis there can be discrepancies between an agent's own beliefs and the beliefs that she ascribes to an actor when she thinks he has some plan. Such discrepancies were associated with the observer's judgement that the actor's plan is invalid. Then the types of any invalidities judged to be present in a plan inferred to underlie a query were shown to affect the content of a cooperative response. I further suggested that, to

[9] Hence, existing PI systems that equate R's and Q's beliefs about actions could, in principle, have handled examples such as (26) which require only the use of (PI1), although they have not done so. Further, while they could have handled the particular type of invalidity that can be inferred using (PI1), without an analysis of the general problem of invalid plans and their effects on cooperative responses, these systems would need to treat this as a special case in which a variant response is required.

guarantee a cooperative response, the observer must attempt to ascribe to the questioner more than just a set of beliefs and intentions sufficient to believe that he has some plan: she must also attempt to ascribe to him beliefs that explain those beliefs and intentions. The *eplan* construct was introduced to capture this requirement. Finally, I described the process of inferring eplans—that is, of ascribing to another agent beliefs and intentions that explain his query and can influence a response to it.

REFERENCES

[1] James F. Allen. *A Plan Based Approach to Speech Act Recognition.* Technical Report TR 121/79, University of Toronto, 1979.

[2] James F. Allen. Recognizing intentions from natural language utterances. In Michael Brady and Robert C. Berwick, editors, *Computational Models of Discourse,* pages 107-166, MIT Press, Cambridge, Mass., 1983.

[3] James F. Allen. Towards a general theory of action and time. *Artificial Intelligence,* 23(2):123-154, 1984.

[4] Michael Bratman. *Intention, Plans and Practical Reason.* Harvard University Press, Cambridge, Ma., forthcoming.

[5] Michael Bratman. Taking plans seriously. *Social Theory and Practice,* 9:271-287, 1983.

[6] M. Sandra Carberry. *Pragmatic Modeling in Information System Interfaces.* PhD thesis, University of Delaware, 1985.

[7] Philip R. Cohen and Hector J. Levesque. Speech acts and rationality. In *Proceedings of the 23rd Conference of the Association for Computational Linguistics,* pages 49-59, Stanford, Ca., 1985.

[8] R. E. Fikes and Nils J. Nilsson. Strips: a new approach to the application of theorem proving to problem solving. *Artificial Intelligence,* 2:189-208, 1971.

[9] Alvin I. Goldman. *A Theory of Human Action.* Prentice-Hall, Englewood Cliffs, N.J., 1970.

[10] Aravind K. Joshi, Bonnie Webber, and Ralph Weischedel. Living up to expectations: computing expert responses. In *Proceedings of the Fourth National Conference on Artificial Intelligence,* pages 169-175, Austin, Tx., 1984.

[11] Diane Litman. *Plan Recognition and Discourse Analysis: An Integrated Approach for Understanding Dialogues.* PhD thesis, University of Rochester, 1985.

[12] Edwin P.D. Pednault. *Preliminary Report on a Theory of Plan Synthesis.* Technical Report 358, SRI International, 1985.

[13] C. Raymond Perrault and James F. Allen. A plan-based analysis of indirect speech acts. *American Journal of Computational Linguistics,* 6:167-182, 1980.

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

[15] Earl D. Sacerdoti. *A Structure for Plans and Behavior.* American Elsevier, New York, 1977.

[16] Candace L. Sidner. Plan parsing for intended response recognition in discourse. *Computational Intelligence,* 1(1), 1985.

[17] Candace L. Sidner. What the speaker means: the recognition of speakers' plans in discourse. *International Journal of Computers and Mathematics,* 9:71-82, 1983.

[18] David E. Wilkins. Domain-independent planning: representation and plan generation. *Artificial Intelligence,* 22:269-301, 1984.