Intentions in Communication.
Cohen, Morgan + Pollack.
MIT Press, 1990

Chapter 5

Plans as Complex Mental Attitudes

Martha E. Pollack

1 Introduction

There are plans and there are plans. There are the plans that an agent "knows": essentially recipes for performing particular actions or for achieving particular goal states. And there are the plans that an agent adopts and that subsequently guide his action. The distinction is between knowing that a plan for assassinating the president is shooting him, and actually planning to assassinate the president by shooting him.

To keep matters straight, we can refer to what one knows when he knows a way to do something as a *recipe-for-action*. Schematically, if we let r denote some recipe-for-action, A denote an agent, and PLANS be a binary relation between agents and the recipes-for-action they adopt, PLANS(A, r) will denote that A has a plan to do r. (I am temporarily suppressing issues of time.) A potential ambiguity remains, however. For there is A's plan r, the particular recipe-for-action he has adopted, and there is the state of mind that A is in when PLANS(A, r) is true. It is this latter distinction that Bratman is addressing when he notes that in speaking of an agent's plan we might "mean an appropriate abstract structure—some sort of partial function from circumstances to actions, perhaps. On the other hand, [we might] mean an appropriate state of mind, one naturally describable in terms of such structures" (1987, 271). We thus need more terminology: if PLANS(A, r) is true, we can say that the recipe-for-action r is the *object* of A's plan, which is itself what Bratman calls a "state of mind." Indeed, I will argue that the plan itself is usefully seen as a complex mental attitude, one comprising a structured collection of beliefs and intentions.

Artificial intelligence (AI) research in plan generation has always been primarily concerned with the recipes-for-action that are the object of an

Preparation of this paper was supported by a gift from the System Development Foundation. The research was done as part of my Doctoral thesis (Pollack 1986), which was supported by a gift from the System Development Foundation, by an IBM Graduate Fellowship, by the Defense Advanced Research Projects Agency under contract N000039-84-K-0078, and by the Office of Naval Research under contract N00014-855-C-0013. My sincere thanks to Barbara Grosz and to all the others, too numerous to list here, who contributed to my thesis effort.

agent's plans. The focus of such research has been on automating the process by which an agent can compute a recipe-for-action. Thus, Nilsson describes plan generation as "the problem of synthesizing a sequence of robot actions that will (if properly executed) achieve some stated goal, given some initial situation" (1980, 275). Although the assumption has always been that the computed recipe-for-action would be adopted as the object of a plan, and some robots have been designed that did actually adopt and execute the recipe-for-action, the plan generation literature has paid very little attention to plans qua mental attitudes.[1] The same is true of much of the plan inference literature in AI, which began later than the work on plan generation and which inherited many of its techniques.

In this paper I claim that a model of plan inference adequate to support a theory of cooperative communication must concern itself with the structure of the complex mental attitude of having a plan, as well as with the structure of the objects of that attitude. In section 2 I recount the traditional AI approach to performing plan inference in communication and describe some limitations of that approach. In section 3 I present an alternative analysis of plans, one that emphasizes their nature as action-guiding, complex mental phenomena. In section 4 I describes the role of this analysis of plans in a theory of cooperative communication that is not subject to the limitations discussed in section 2. In section 5 I return to the traditional AI model of plans and show how it is subsumed by the model presented here.

The theory of cooperative communication discussed in section 4 has been developed in some detail and has been implemented in SPIRIT, a system that reasons about the plans underlying the queries that it is asked and that generates appropriate responses. SPIRIT is able to infer plans even when they are invalid, that is, when their objects are not recipes-for-action that SPIRIT would itself construct. This ability distinguishes SPIRIT from earlier plan inference systems, which do not rely on a careful analysis of plans as complex mental attitudes. Details of SPIRIT and of the plan inference framework it embodies can be found in Pollack 1986.

2 *The Traditional Approach*

If you overheard the following conversation, you would probably find it quite unremarkable:

> *A*: "I want to talk to Kathy, so I need to find out the phone number of St. Eligius."
>
> *S*: "St. Eligius closed last month. Kathy was at Boston General, but she's already been discharged. You can call her at home. Her number is 555-1238."

Intuitively, it is quite clear what is occurring in this discourse. Agent *A* believes that Kathy is at St. Eligius and plans to call her there. Agent *S* believes that *A*'s intended act of calling St. Eligius cannot be performed, since St. Eligius is closed. Moreover, *S* believes that even if *A* could call St. Eligius, it would not contribute to his goal of talking to Kathy, because she is actually at home. Consequently, *S*, being cooperative, provides *A* with information she believes will contribute to his goal: she tells him the correct phone number for reaching Kathy. She also tells him why she believes the information he requested was not appropriate to his goal.

Conversations such as this one, in which the beliefs of the inferring agent differ significantly from the beliefs of the actor whose plan she is inferring, provide a serious challenge to most existing AI systems for plan inference in communication (see, for instance, Allen 1979; Carberry 1985; Litman 1985; Sidner 1985). To see why, it is instructive to consider briefly how these systems work. Each of them has an *operator library*, which encodes a set of recipes-for-action. These operators are a direct outgrowth of the representations first developed in the STRIPS system (Fikes and Nilsson 1971) and later expanded in the NOAH system (Sacerdoti 1977). Each operator may contain some or all of the following parts:

- a *header*, which names the action α for which the operator is a recipe;
- a *precondition list*, which describes what must be true for α to be performed;
- an *effect list*, which describes what will be true after α is performed;
- a *list of constraints*, which describes restrictions on legal instantiations of the operator;[2]
- a *body*, which may be a set of subactions whose performance constitutes performance of α, or a set of subgoals whose achievement constitutes performance of α.

The operators in the library represent relatively simple recipes-for-action. More-complex recipes can be constructed out of the simpler ones. Both in plan generation and in plan inference, the more-complex recipes-for-action are represented as directed acyclic graphs, whose nodes are labeled with operator headers and propositions; when a node is labeled with a proposition *P*, it should be interpreted as denoting any action that would achieve *P*. Graphs, rather than linear orderings, are necessary because the recipes are seen as having both a hierarchical and a temporal dimension. Recipe graphs are sometimes referred to as macro-operators.

All the plan inference systems include a set of rules for constructing recipe graphs out of the simpler recipes in the operator library. Each rule states the conditions under which a piece of a recipe graph can be constructed. The conditions for constructing a subgraph always refer both to the subgraphs that have already been constructed and to operators in the operator library. A typical rule, for example, allows a subgraph that

A plan subgraph with nodes α and β and an arc from α to β,

$$\begin{array}{c}\beta\\ \uparrow\\ \alpha\end{array}$$

can be constructed provided either α or β is already in the subgraph and $\alpha R \beta$ holds, where R is one of the following relations:

1. R = *causes*; that is, β is on the effect list of α, so there is an operator in the operator library of the form

   Header: α
   Effects: ... β ...

2. R = *is-a-precondition-of*; that is, α is on the precondition list of β, so there is an operator in the operator library of the form

   Header: β
   Preconditions: ... α ...

3. R = *is-a-way-to*; that is, α is part of the body of β, so there is an operator in the operator library of the form

   Header: β
   Body: ... α ...

Figure 5.1
Rules for constructing plan graphs.

includes a node labeled α to be expanded by adding an arc from α to a new node labeled β, if there is an operator whose header is β and which includes α in its body. In all, there are only three basic conditions under which the various plan inference systems will construct a recipe subgraph. These are shown in figure 5.1. Note that each condition corresponds to one of the ways that actions and propositions can be related in an operator.[3]

Plan inference systems begin with some action α believed to be part of the recipe-for-action that is the object of the actor's plan. They then attempt to construct larger recipes-for-action by repeatedly applying inference rules, until one of the computed recipes satisfies some termination condition. (Typically, the termination condition is that the recipe constructed is for an action that an agent in the domain is likely to want to perform.) The computed recipe is taken to be the object of the actor's plan.

In fact, the inference rules are often written in such a way as to acknowledge the importance of plans as mental attitudes; after all, inferring another agent's plan means figuring out what actions he "has in mind." Consider Allen's model, which was one of the earliest accounts of plan inference in

conversation and inspired much of the subsequent work in the field (Allen 1979, 1983). Allen writes AW(*P*), where *P* is some proposition, to express the fact that *A* has a plan to achieve *P*; AW(ACT), where *ACT* names some action, expresses the fact that *A* has a plan to perform *ACT*. A typical plan inference rule, then, is expressed as

SBAW(P) → SBAW(ACT), if P is a precondition of ACT.[4]

This rule corresponds to one direction of application of Condition 1 of figure 5.1.[5] It can be glossed as follows: if the system (inferring agent) believes that the actor wants some proposition *P* to be true, then the system may draw the inference that the actor wants to perform some action *ACT* of which *P* is a precondition. Notice that it is left unstated precisely who it is—the system or the actor—that believes that *P* is a precondition of *ACT*. If we take this to be a belief of the system, it is not clear that the system will infer the actor's plan; but, on the other hand, if we take it to be a belief of the actor, it is unclear how the system 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: these represent recipes-for-action that are assumed to be mutually known to the system and the actor.

In effect, the "SBAW" context is transparent to the reasoning process that is performed by Allen's system: its reasoning is all performed directly on its object, and the B and W operators are carried directly from antecedent to consequent in each inference rule. Without any repercussions, the B and W operators can be omitted, resulting in rules that are completely equivalent to those in figure 5.1. In fact, in his example plan graphs Allen often omits the B and W operators entirely, and this practice has been continued in more recent work in plan inference (Carberry 1985; Litman 1985). For Allen, as well as for most of those whose work he inspired, "the properties of the W operator [are] specified only by the ... plan inference rules." Allen's W operator is equivalent to the PLAN relation introduced here in section 1: it is the relationship that holds between an agent and a recipe-for-action he adopts.[6] His plan inference rules apply essentially directly to the object of the plan under construction. Thus Allen, and his followers, analyze the state of having a plan only in terms of the structure of its object.

In sum, the traditional approach to plan inference has been to reason directly about the object of a actor's plan, constructing a recipe-for-action using a library of simpler recipes that are assumed to be mutually known to the actor and the inferring agent. Though this approach has proved to be successful for modeling a number of conversational phenomena, it has at least three shortcomings that preclude it from handling in a principled way dialogues like the one presented at the beginning of this section—

dialogues in which one of the participants has a plan that the other deems to be invalid.

First, such a system will infer a plan subgraph linking nodes labeled X and Y, where X and Y are either operator headers or propositions, only if X and Y are both encoded in the system's operator library in one of the configurations shown in figure 5.1. This fact, combined with the fact that the operator library includes only valid domain information, may at first seem to suggest that traditional systems cannot infer *any* invalid plans. It turns out, however, that there are certain types of invalid plans that can, in principle, be inferred by a traditional system. For example, such a system could use a rule based on Condition 3 of figure 5.1 to infer a subgraph joining some actions α and β, even if the system also believes that one of the preconditions in the operator relating α and β is false.[7] However, without understanding why such a subgraph might be part of the object of the actor's plan, the system will not able to determine whether or not it is reasonable to include such a subgraph in the inferred plan. In fact, existing systems have not focused on the inference of such invalid plans and indeed have included various control heuristics that bias the inference process against finding them (see, for example, Allen's Heuristic H1 (1983, 127)). The first shortcoming of the traditional approach is an inability to state within it general heuristics for rule application.

Second, there are other types of invalid plans that cannot be inferred at all under the traditional approach. For instance, an actor A may be relying upon a simple recipe-for-action that is not in the system's operator library. In the case mentioned in the previous paragraph, S and A mutually know a recipe-for-action linking α and β: A's plan is invalid because the necessary preconditions for doing β by doing α do not hold. However, there are other cases in which A plans to do some action β by doing some α, where S does not at all believe that α will, under any set of preconditions, lead to β. (Indeed, S may even believe that α—or β, or both—cannot be done at all.)

One way to enable a system to reason about plans with such invalidities is to encode directly sets of erroneous beliefs that handle such cases by encoding directly sets of erroneous recipes that their users are likely to have; this is the approach taken by many computer-aided instruction (CAI) systems (Brown and Burton 1978; Collins, Stevens, and Goldin 1979; Genesereth 1979; Woolf and McDonald 1983). Although this seems to be a useful strategy, it is necessarily incomplete. It is impossible for any person to have complete knowledge of the potential beliefs of other people, since the range of beliefs is, in principle, infinite. This means that system designers cannot anticipate a priori all of the potential misconceptions the users of their system may have. It also means that the intelligent agent that such systems emulate—the human being—cannot know a priori all of the potential misconceptions that people asking her questions might have. Sometimes she will have to deal with a novel (to her) belief. A more general approach is to try to understand the structure of the beliefs that an agent has when she has some plan, thereby making it possible to apply various belief-attribution techniques to the problem of plan inference.

A third weakness of the traditional approach concerns the generation of appropriate responses in conversation. To explain why S's response in the example dialogue is a cooperative one, it is necessary to understand what beliefs S is attributing to A by virtue of attributing to her some plan.

To handle these issues—to make room in the theory of plan inference for discrepancies between the beliefs of the actor and the inferring agent, thereby allowing for the proper treatment of "invalid" plans—it is necessary to undertake an analysis of what the context "SBAW" means: to analyze plans as mental phenomena. We now turn to that task.

3 *An Alternative Model of Plans*

To develop an account of plans as mental phenomena, we could begin with the traditional AI models, which, as we have seen, have focused on the objects of plans: we could attempt to determine what sort of mental states would reasonably have such objects. However, it will prove to be more fruitful to begin with our commonsense conceptions of what it means to have a plan. The resulting model can be shown to subsume and make more precise the model implicit in the traditional AI accounts.

3.1 *The Belief Component of Plans*

Let us begin with the plan (referred to earlier) to ask Kathy how she is feeling. I plan to do this, believing that Kathy is at St. Eligius, by finding out the phone number of St. Eligius, calling there, and then saying to Kathy "How are you doing?"[8] The performance of these acts is meant to "entail"—in a sense of *entail* yet to be further specified—the performance of my goal act.

One's plans, however, may fail. If, unbeknownst to me, Kathy has already gone home, then my plan will not lead to my goal of asking her how she is feeling. For me to have a plan to do β, which consists in the doing of some collection of acts 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.[9] This insight is at the core of a view of plans as mental phenomena; on this view, plans "exist"—that is, gain their status as plans—by virtue of the beliefs of the person whose plans they are.

So far, then, I have associated the state of having a plan to do β with a belief that executing some collection of acts Π will lead to doing β. Note that the temporal ordering of the acts is an essential part of the plan: I may well have a plan to prepare onions for a sauce by chopping them and then

sautéing them—believing that by so doing I will perform my goal—and not have a plan that involves sautéing and then chopping the onions. However, the ordering need not be total: as Sacerdoti (1977) demonstrated, there are many plans with objects that include acts whose temporal order with respect to one another is irrelevant. For example, I may have a plan to set the table that includes the following acts: carrying the flatware, plates, and glasses to the table; setting out the flatware; setting out the plates; and setting out the glasses. I may believe that is essential that the first act in the list be performed prior to the others, but I may also believe that the other three acts can be performed in any order with respect to one another and can even be interleaved. Of course, when I actually execute my plan, the acts I perform will be totally ordered with respect to one another. So there is a sense in which the beliefs that are part of my plan are partial.

There is also another sense in which my beliefs may be partial: they may concern acts only to an arbitrary level of abstraction. The set of acts that I believe will entail my asking Kathy how she is feeling includes the act of finding out the phone number of St. Eligius. It includes this despite the fact that I may not yet have considered how I will do this—for example, whether I will call the information operator or look in a phone book.[10]

An agent A's belief that performing the acts in Π will entail performing β is not by itself sufficient to guarantee that A has a plan, consisting of doing Π, to do β. To see why not, consider the following scenario. Suppose that I decide that while I am finding out the phone number of St. Eligius (say, by looking in the phone book), I might as well at the same time find out the phone number of my bank (perhaps because I know I have to call there later to check on a wire transfer). As before, I believe that finding out and dialing the phone number of St. Eligius, and then saying certain words, will entail my asking Kathy how she is feeling. I do not believe that this will cease to be true if I also find out my bank's phone number. Thus, there is a temporally ordered collection of acts Π, which equals finding out the phone number of St. Eligius, finding out the phone number of my bank, calling St. Eligius, and saying to Kathy "How are you doing?", such that I believe that executing Π will lead to my goal of asking Kathy how she is feeling.[11] However, it seems incorrect to say that my plan to ask Kathy how she is feeling includes my act of finding out the phone number of my bank; instead, this act is part of another plan (to check on my wire transfer) that I intend to interleave with my original plan. For an act to be included in my plan, I must believe that it plays a role in that plan; for the case at hand I do not believe that finding out the phone number of the bank plays any role in my plan to ask Kathy how she is feeling.

3.2 *Playing a Role in a Plan*

What does it mean for an act to play a role in a plan? Consider once more my plan to ask Kathy how she is feeling. Part of this plan, I claimed,

involves my calling St. Eligius. What would we say about my beliefs about these two acts: my asking Kathy how she is feeling and my calling St. Eligius? We might say that I believe that the latter will enable the former—that is, that by calling St. Eligius, I will establish a communication channel (a phone link) to Kathy, which will enable my saying something to her. Similarly, we might say that I believe that finding out the phone number of St. Eligius plays a role in my plan because I believe that doing the former will enable my calling St. Eligius, which itself plays a role in my plan. In general, then, if an agent believes that doing one act α will enable either his goal or some other act γ that plays a role in his plan, then α may play a role in his plan. In order to strengthen the "may" to "will," we need to consider the agent's intentions. I will discuss this presently.

Next consider the relationship between my acts of saying to Kathy "How are you doing?" and asking her how she is feeling. We would not say that I believe that the first act will enable the second. Instead, we could describe my beliefs using the "*by*-locution" in English: we could say that I believe that *by* saying "How are you doing?", I will be asking Kathy how she is feeling. Similarly, consider a slightly more detailed analysis of my plan to ask Kathy how she is feeling, which makes explicit the way in which I go about finding out the phone number of St. Eligius. One way in which I might plan to do this is by looking the number up in the phone book. We would not say that I believe that my looking up St. Eligius's phone number will *enable* my finding it out; rather, we would say that I believe that by looking it up, I will find it out. Or I might, instead of looking up the phone number, plan to discover it by getting my office mate to tell it to me, as a result of my asking her to tell it to me, which consists in my uttering the question "Do you know the phone number of St. Eligius?" Once again, we would not say that I believe that my uttering the question "Do you know the phone number of St. Eligius?" will *enable* my asking my office mate to tell me the phone number of St. Eligius, nor would we say that my asking her to tell me the phone number will *enable* my getting her to tell it to me. What we would say is that I believe that *by* uttering the question, I will be asking my office mate to tell me the phone number; that *by* asking her, I will be getting her to tell it to me; and that *by* getting her to tell it to me, I will be finding it out. Notice that the *by*-locution does *not* completely correlate with causation: although it is true that asking to be told the phone number seems to relate causally to getting a listener to tell you the phone number, there is no such causal flavor to the relation between uttering the question and asking to be told the phone number.

The claim is that if an agent believes that by doing α, he will be doing either his goal act β or some other act γ that plays a role in his plan, then α may play a role in his plan. Of course, to some extent this just begs the

question, since I have left vague the conditions under which one act can be said to be done *by* doing another. Goldman's (1970) concept of *generation* can be used to make these conditions more precise (Pollack 1986, 52–67). I will adopt the term *generation* to describe the relation between two acts that we commonly express with the *by*-locution. What is important to notice here is that there does seem to be an intuitive difference between the relation between acts that we describe as enablement and the one I am now calling generation. Just as we would not describe the relations between the acts discussed in the previous paragraph as enablement, so we would not, in general, use *by* to describe the relations earlier discussed as examples of enablement. We would not, for example, say that I believe that I can call St. Eligius by finding out the phone number there.[12]

What does the difference between enablement and generation consist in? Most importantly, it is the case that when one action α generates another action β, then the agent need only do α and β will automatically be done also. However, when α enables β, then the agent needs to do something more than α to guarantee that β will be done. I cannot simply find out the phone number of St. Eligius and then rationally expect that I will have called St. Eligius. But if I utter the words "Do you know the phone number of St. Eligius?" in the appropriate circumstances, then I need do nothing more to have asked my office mate the phone number of St. Eligius. And having done that, if the circumstances are right (say, my office mate knows the phone number of St. Eligius and is willing to tell it to me), then I need do nothing more to have caused my office mate to tell me St. Eligius's phone number.

Of course, if the circumstances are not right, then by asking my office mate for the phone number, I will not have caused her to tell it to me. For instance, she might not hear me ask her. In that case I may repeat my question. It is important not to confuse things here. It is true that in this case I *do* "do something more" than my original act of asking for the phone number in order to perform the act of causing my office mate to tell it to me. But here my first act—of asking for the phone number—does *not* generate my act of causing my office mate to tell it to me. My second act—repeating my request—might do so, if my office mate hears me this time and responds, but then I have not "done anything more" than this second act, of repeating my question, to cause my office mate to tell it to me.

Analogously, I might dial the phone number and get a busy signal. I then need to do something more to establish a communication channel to Kathy. I might dial the number again, or I might drive to St. Eligius. In either case, if I succeed in establishing a communication channel, I succeed in doing so by dialing the second time, or by going to St. Eligius: it is my

act of *re*dialing or my act of going to St. Eligius, and not my original act of dialing, that generates my act of establishing the communication channel.

3.3 *The Intention Component of Plans*

So far we have seen that for an agent A to have a plan to do β that consists of doing Π, he must have a certain set of beliefs about the acts that Π comprises. Specifically,

> 1. A must believe that executing the acts in Π, in their (possibly partial) temporal order, will entail his performance of β
>
> and
>
> 2. A must believe that each act α in Π plays a role in his plan; that is, either he believes that by doing α he will do β or some other γ that plays a role in his plan (in other words, that α generates β or γ), or he believes that doing α will enable doing β or some other γ that plays a role in his plan.

Condition 1 can be thought of as a sufficiency condition: it guarantees that if A believes that α is part of what he will do to achieve his goal, then α is in the plan. Condition 2 can be thought of as a necessity condition: it guarantees that if α is part of A's plan, he believes it will play a role in achieving his goal.

Although these beliefs are necessary, they are not sufficient to guarantee that doing Π is A's plan to do β. It is also necessary that A have a certain set of intentions with respect to Π. In particular, for my plan to do β to consist in my doing Π, I must intend to execute each of the acts in Π. Let Π once more be finding out the phone number of St. Eligius, calling St. Eligius, and saying "How are you doing?" I may believe that executing Π would entail my asking Kathy how she is feeling, but if I intend instead to wait until next Thursday and then talk to her face to face, Π is not a plan I have to ask Kathy how she is feeling.

Further, in order for doing Π to count as my plan to β, not only must I intend to execute Π, but I must also intend it *as a way* of doing β; that is, I must intend to do the acts in Π in order to do β. Imagine that I am a teenager whose parents have forbidden me to stay out past midnight. Imagine further that there is some club, "The May Day Late-Night Club," whose membership is limited to those who show up at the movies at 1 a.m. on May 1. Now let Π be the singleton set going to the movies at 1 a.m. on May 1. I may intend to execute Π, and I may believe that doing so will entail both becoming a member of the May Day Late-Night Club and aggravating my parents. But if I intend my execution of Π as a way to do the former—and *not the latter*—then doing Π will count as a plan I have to join the Late-Night Club, but not as a plan I have to aggravate my parents.

Notice the this is true even if I intend to aggravate my parents in some other way—say, by getting a Mohawk haircut.

An argument similar to that made in the belief case also applies here. Thus, not only must I intend to execute Π in order to do β, I must also intend each act α in Π to play a role in my doing β; that is, I must intend each α either to generate or to enable β or some other γ that itself plays a role in my plan. Thus, we can state three more conditions on A's having a plan to do β that consists in doing Π, namely:

3. A must intend to execute each act α in Π, in the (possibly partial) temporal order;
4. A must intend to execute Π as a way of doing β;

and

5. A must intend each act α in Π to play a role in his plan. In other words, either he must intend by doing α to do β or some other γ that plays a role in his plan, or he must intend by doing α to enable doing β or some other γ that plays a role in his plan.

The close parallel between Conditions 2 and 5 should lead us to ask whether one subsumes the other. The discussion so far has shown that having the beliefs defined in Condition 2 does not entail having the intentions in Condition 5. But does having the intentions in Condition 5 entail the beliefs in Condition 2?

The question of whether an agent's intention to do α entails a belief that the agent will do α has been debated in the philosophical literature, and no consensus seems to have been reached. Whereas Grice (1971), for one, answers in the affirmative, Davidson (1980) goes to great lengths to provide a counterargument.[13] A telling comment on the controversy is provided by Bratman (1983), who notes that "plans normally support expectations of their successful execution..., [although] there may still be cases in which I plan to A but do not believe I will" (p. 286, fn. 4). Within this paper it will prove to be sufficient to assume this "normal" state of affairs and accept the view that if one intends to X, one must believe that one will.

Given this, we can see that Condition 5 directly entails Condition 2. For if, for each α in Π, A intends α to play a role in his plan, then A also believes that α will play a role in his plan. If A intends to do β or some other γ by doing α, he also believes that he will do β or this other γ by doing α; if he intends to enable β or some other γ by doing α, he also believes that he will enable β or this other γ by doing α.

Similarly, Condition 4 entails Condition 1. For if A intends to do Π as a way of doing β, then A must believe he will do β by doing Π, and this is exactly Condition 1.

To state the commonsense requirements on A's having a plan, it is thus sufficient to state the conditions on intending—Conditions 3 through 5. However, although the conditions on belief—Conditions 1 and 2—are entailed by Conditions 4 and 5, respectively, and are thus redundant, it is worth keeping them in mind, for it will turn out that when an inferring agent deems an actor's plan invalid, it is because she believes it includes invalid beliefs. In fact, it is even worthwhile to make explicit yet another belief that is requisite to having a plan. Condition 3 asserts that A intends to execute each act α in Π, this then entails that A believes he will do each act α in Π. And this belief, in turn, entails a belief that A can do each act α in Π. Adding this condition to the definition, we can summarize the analysis of "having a plan" as follows:

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

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

4 *Plans in Cooperative Communication*

We can now consider how the model of plans just developed can be incorporated in a theory of communication that does not suffer from the shortcomings described in section 2—that is, one in which even invalid plans can be reasoned about properly. To begin, it is useful to develop a more formal representation of the analysis given in Definition P0.

4.1 A Representation for Simple Plans

I will restrict my attention to a subset of plans, which I 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—in other words, if it includes no acts that the agent believes are related to one another by enablement. The representation language I will use builds upon Allen's (1984) interval-based temporal logic, a typed first-order predicate calculus. In particular, I will make use of the predicates HOLDS and OCCURS. HOLDS is a binary relation over propositions and time intervals: HOLDS(P, t) is taken to be true if and only if proposition P holds throughout time interval t. OCCURS, as I will use it, is a ternary relation: OCCURS(α, A, t) is true if and only if agent A performs an act of type α during time interval t.

Note that the first argument to OCCURS is an act-type. The distinction between act-types and actions is crucial. Actions or acts—I will use the two terms interchangeably—can be thought of as triples of act-type, agent, and time. Thus, typing **DEL .** is an example of an act-type, whereas my typing **DEL .** right now is an example of an action. Generation is a relation over actions, not over act-types. Sometimes an act of typing **DEL .** will generate an act of deleting the current mail message (say, when the former act is performed while the agent is using a particular electronic mail system). But not every case of an agent's typing **DEL .** will result in the agent's deleting the current message; for example, my typing it just now did not, because I was not typing it to a computer mail system. However, when an act of A's doing α at time t generates an act of his doing β at time t, there are certain conditions C such that any time there occurs an act of α while C holds, there will also occur a simultaneous act of β. The regularity of the generation-enabling conditions C is what enables us to reason about whether by doing some action we will do another action, and consequently what enables us to construct and reason about (simple) plans.

To encode these regularities, I will introduce the following abbreviatory device: I will say that act-type α *conditionally generates* act-type β under conditions C, and I will write CGEN(α, β, C), where C denotes the generation-enabling conditions relating α and β. Thus, the CGEN predicate is defined as follows:

Definition C1
$CGEN(\alpha, \beta, C) \leftrightarrow$

1. $\forall A \forall t_1[[HOLDS(C, t_1) \wedge OCCURS(\alpha, A, t_1)]$
   $\rightarrow OCCURS(\beta, A, t_1)] \wedge$
2. $\exists A \exists t_2[OCCURS(\alpha, A, t_2) \wedge \neg OCCURS(\beta, A, t_2)] \wedge$
3. $\exists t_3 \exists A[HOLDS(C, t_3) \wedge \neg OCCURS(\beta, A, t_3)]$.

It is then straightforward to define the generates relation, GEN, in terms of CGEN:[14]

Definition G1
$GEN(\alpha, \beta, A, t) \leftrightarrow \exists C[CGEN(\alpha, \beta, C) \wedge HOLDS(C, t)]$.

That is, agent A's doing α at time t will generate his doing β if and only if there are some generation-enabling conditions relating α and β and, further, those conditions hold at time t.

In translating Definition P0 into the representation language, I will also make use of several other relations, which I will treat as primitive in this paper. (See, however, Pollack 1986, 67–72.) The relation EXEC(α, A, t) will be taken to be true if and only if the act of A's doing α during time interval t is *executable*. The relation BEL(A, P, t) is true if and only if agent A believes

proposition P throughout time interval t; INT(A, α, t_2, t_1) is true if and only if throughout time t_1, A intends to do α at time t_2. I also make use of one function: *by* maps two act-types into a third, composite act-type—$by(\alpha, \beta)$ denotes the act of doing β by doing α.

These various components of the representation language then can be combined in Definition P1, which encodes in the representation language the definition of having a simple plan:

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

1. $BEL(A, EXEC(\alpha_i, A, t_2), t_1)$, for i = 1, . . . , n $\wedge$
2. $BEL(A, GEN(\alpha_i, \alpha_{i+1}, A, t_2), t_1)$, for i = 1, . . . , n − 1 $\wedge$
3. $INT(A, \alpha_i, t_2, t_1)$, for i = 1, . . . , n $\wedge$
4. $INT(A, by(\alpha_i, \alpha_{i+1}), t_2, t_1)$, for i = 1, . . . , n − 1.

The left-hand side of Definition P1 denotes that the agent A 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 Definition P1 corresponds directly to Definition P0, except that, in keeping with the restriction to simple plans, specific assertions about each act generating another replace the more general statement regarding the fact that each act plays a role in the plan. Clause 1 of Definition P1 captures Clause 1 of Definition P0.[15] Clause 2 of Definition P1 captures both Clauses 2 and 3 of Definition P0: when i takes the value $n - 1$, Clause 2 of Definition P1 captures the requirement, stated in Clause 2 of Definition P0, that A believes his acts will entail his goal; when i takes values between 1 and $n - 2$, it captures the requirement of Clause 3 of Definition P0, that A believes each of his acts plays a role in his plan. Similarly, Clause 3 of Definition P1 captures Clause 4 of Definition P0, and Clause 4 of Definition P1 captures Clauses 5 and 6 of Definition P0.

4.2 Invalid Plans and Cooperative Responses

Given Definition P1, it is straightforward to state what it means for an agent to have an invalid simple plan: A has an invalid simple plan if and only if he has the set of beliefs and intentions listed in Definition P1, where one or more of those beliefs is incorrect, and, consequently, one or more of the intentions is unrealizable. We can also see what it means for one agent to believe that another has a simple plan: I will say that S believes that A has some (simple) plan to do the action β by doing the actions $\alpha_1, \ldots, \alpha_n$ if S believes that A has the configuration of beliefs and intentions represented in Definition P1. And these two statements can be combined, so that S will be said to believe that A has a simple plan that is invalid if S believes to be

false some belief she attributes to A virtue of attributing to him some simple plan (and consequently, S also believes to be unrealizable some intention she attributes to A in virtue of attributing to him that same simple plan).

The structure of Definition P1 suggests that there are two types of plan invalidities, corresponding to the two types of beliefs that are part of the mental attitude of having a simple plan. An incorrect belief corresponding to Clause 1 of Definition P1 will indicate a plan that includes an intention to do an *unexecutable act*, and an incorrect belief corresponding to Clause 2 will indicate what I will call an *ill-formed plan*: one in which the intended acts will not lead to the goal. Of course, in accounting for cooperative conversation, what is at issue is not the absolute correctness of the actor's beliefs but, as explained above, the inferring agent's beliefs about the beliefs she attributes to the actor. So, I will say that S judges A's plan to contain an unexecutable act if S believes to be false one of the beliefs she attributes to A in satisfaction of Clause 1 of Definition P1. Likewise, I will say that S judges A's plan to be ill formed if she believes to be false one of the beliefs she attributes to A in satisfaction of Clause 2 of Definition P1. Of course, this is not to say that any agent S would assent to a description of A's plan as "ill formed": to say that an agent believes a plan is ill formed is to describe a configuration of beliefs that agent has.

Consider again the dialogue presented in section 2:

> A: "I want to talk to Kathy, so I need to find out the phone number of St. Eligius."
>
> S: "St. Eligius closed last month. Kathy was at Boston General, but she's already been discharged. You can call her at home. Her number is 555-1238."

We can accout for S's response there by assuming that she believes that A has a certain set of beliefs and intentions satisfying Definition P1. A portion of those beliefs is shown in figure 5.2. S believes that A believes that calling St. Eligius at time t_2 is executable, though S believes it is not executable, and she informs A of this in her response. Also, S believes that A believes that the act of calling St. Eligius will generate the act of establishing a communication channel to Kathy. S believes that it will not—that even if A could call St. Eligius, that act would not have the desired effect. This belief also affects S's response.

$$BEL(S, BEL(A, EXEC(call(St.\ Eligius), A, t_2), t_1), t_1)$$
$$BEL(S, BEL(A, EXEC(establish\text{-}channel(Kathy), A, t_2), t_1), t_1)$$
$$BEL(S, BEL(A, GEN(call(St.\ Eligius), establish\text{-}channel(Kathy), A, t_2), t_1), t_1)$$
$$BEL(S, INT(A, call(St.\ Eligius), t_2, t_1), t_1)$$
$$BEL(S, INT(A, establish\text{-}channel(Kathy), t_2, t_1), t_1)$$
$$BEL(S, INT(A, by(St.\ Eligius), establish\text{-}channel(Kathy), t_2, t_1), t_1)$$

Figure 5.2
The plan S infers for A.

Strategies for producing cooperative responses to questions must incorporate decisions about what information to include. The view of plans developed here—in which plans are seen as complex mental attitudes, and plan inference is seen as the process of attributing such attitudes to an actor—provides the basis for determining a class of information that may need to be included in a cooperative response to a question. Specifically, to be cooperative, a response may need to include information about the particular discrepancies S finds between her own beliefs and those she attributes to A as part of her belief that he has some particular plan. Note, however, that a cooperative response may not necessarily include all such information, for S may deem some or all of it to be irrelevant. The plan inferred to underlie a query and any invalidities it is judged to have are but two factors affecting the response generation process, the most significant others being relevance and salience.

4.3 Explanatory Plans

When S judges A's plan to be invalid in one of the ways discussed above, she has intuitively "made sense" of the plan and understands the source of the invalidities. However, there are also cases in which an inferring agent simply cannot make sense of an actor's query. As a somewhat whimsical example, imagine A saying,

> A: "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?" A's query is *incoherent*: a listener S may be unable to understand why A 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 A'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 A's query would would be coherent but instead to realize that there are many circumstances in which it would not.

The model of plans as I have so far presented it does not distinguish between a query of this type and one in which the inferred underlying plan is ill formed. The reason is that, given a reasonable account of semantic interpretation, it is transparent from the query just given above that A 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 doing the former. Further, as consequences of these intentions A 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 doing the former.[16] But these beliefs and intentions are precisely what are required by Definition P0 to have a plan; and if S could determine that the intended role of the supporting act of standing on his head was generation, then these beliefs and intentions would also be exactly what is required by Definition P1. Consequently, after hearing the query, S can in fact infer a plan underlying A's query, namely, the obvious one: to find out how to stand on his head in order to talk to Kathy. Then, since S does not herself believe that the former act will lead to the latter, on the analysis so far given, we would regard S as judging A's plan to be ill formed. Unfortunately, this is not the desired analysis: the model should instead capture the fact that S cannot make sense of A's query here—that it is *incoherent*.

To capture the difference between ill-formedness and incoherence, I will claim that, when an agent S is asked a question by an actor A, S needs to attempt to ascribe to A more than just a set of beliefs and intentions satisfying Definition P1. Specifically, for each belief satisfying Clause 2 of Definition P1, S must also ascribe to A another belief that explains the former in a certain specifiable way. The beliefs that satisfy Clause 2 are beliefs about the relation between two particular actions. For instance, the plan underlying the example query includes A's belief that his action of calling St. Eligius 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 A has about the relation between the act-types "calling a location" and "establishing a communication channel to an agent." A 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 CGEN relation introduced earlier. So, for instance, S may attribute to A a belief that we can express as follows:

$$BEL(A, CGEN(call(X), establish\text{-}channel(Y), at(X, Y)), t_1).$$

This belief, combined with a belief that Kathy will be at St. Eligius at time t_2, expains A's belief that, by calling St. Eligius at t_2, he will establish a communication channel to Kathy. In contrast, S may have no basis for ascribing to A beliefs that will explain why he thinks that standing on his head will lead to talking with Kathy. Consequently, she will deem the second example query to be incoherent.

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

Definition P2

$BEL(S, EPLAN(A, \alpha_n, [\alpha_1, \ldots, \alpha_{n-1}], [\rho_1, \ldots, \rho_{n-1}], t_2, t_1), t_1) \leftrightarrow$

1. $BEL(S, BEL(A, EXEC(\alpha_i, A, t_2), t_1), t_1)$, for $i = 1, \ldots, n \wedge$
2. $BEL(S, BEL(A, GEN(\alpha_i, \alpha_{i+1}, A, t_2), t_1), t_1)$, for $i = 1, \ldots, n-1 \wedge$
3. $BEL(S, INT(A, \alpha_i, t_2, t_1), t_1)$, for $i = 1, \ldots, n \wedge$
4. $BEL(S, INT(A, by(\alpha_i, \alpha_{i+1}), t_2, t_1), t_1)$, for $i = 1, \ldots, n-1 \wedge$
5. $BEL(S, BEL(A, \rho_i, t_1), t_1)$, where each ρ_i is $CGEN(\alpha_i, \alpha_{i+1}, C_i) \wedge HOLDS(C_i, t_2)$.

Clauses 1–4 of Definition P2 are similar to Clauses 1–4 of Definition P1. The key clause in the definition of eplans is Clause 5: for S to believe that A has some eplan, she must attribute to A beliefs that explain the other beliefs that are constituents in his plan. I will call the beliefs that S attributes to A in satisfaction of Clause 5 *explanatory beliefs*.

It is now possible to sketch the process of inferring the eplan that underlies a query. To begin, S will believe that A intends to do some act α, possibly because A tells her this in the query. S thus may believe that A has a trivial eplan, that is,

$BEL(S, EPLAN(A, \alpha, [\ ], \rho, t_2, t_1), t_1)$, where ρ is nil (logically true).

Let us suppose then that S has reason to believe—on the basis of something other than A's query itself—that it is plausible that A believes that act-type α conditionally generates some other act-type γ, under some specific condition C, and that she has no reason to suppose that A believes that C will not hold at the intended performance time of his plan. Then S can decide that it is plausible for A to believe that by his act of α, he will do γ and, further, that it is plausible for him to intend to do α in order to do γ, and to intend to do γ. The plausibility of A's having these intentions depends upon the plausibility of his having the aforementioned beliefs, so S will attribute to A as a bundle the plausible intentions and supporting beliefs. That is, S will reason from the trivial eplan to a larger one. The process of belief and intention ascription can be iterated: believing that it is plausible that A intends to do γ, S can then reason about what A might believe he can do by this act. S can also reason about what A might plausibly intend to do in order to do his goal act β and can then iterate as well in the "backward" direction. Again, the reasoning is from one plausible eplan to another.

What is crucial in this picture is the way in which S attributes to A particular explanatory beliefs. As noted, when S decides that it is plausible for A to believe that by doing α he will be doing some γ, the plausibility of that belief must be established on the basis of something other than A's

query itself; otherwise, the very distinction that eplans were set up to support—the distinction between incoherent and ill-formed plans—will collapse. Though it is true that if A says "I want to talk to Kathy, so I need to find out how to stand on my head," S can figure out that A believes that by standing on his head he can talk to Kathy (or at least enable his talking to Kathy), this belief can only be attributed to A on the basis of his query. The types of evidence that can be used to attribute explanatory beliefs to an agent are encoded in a set of *plan inference rules* that describe the reasoning that is permitted in going from one plausible eplan to another. If the plan inference process proceeds only through the use of plan inference rules, the distinction between ill-formed plans and incoherent queries is maintained.

In the simplest plan inference rule, S attributes to A a belief that she herself has:[17]

Rule PI1

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

This rule says that, if S's belief that A has some eplan includes a belief that A intends to do an act α_n, and S also believes that act-type α_n conditionally generates some γ under condition C, then S may infer that A has the additional intention of doing α_n in order to do γ—that is, that he intends to do $by(\alpha_n, \gamma)$. A'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.[18]

One way to view Rule PI1 is as an explication of condition 3 of figure 5.1 in the language of plans as mental phenomena. If a system made use only of Rule PI1 (and its symmetric partner), it would be implicitly committing to the assumption that A has the same beliefs about conditional generation as it does. Additional plan inference rules are needed to move beyond this assumption. In one obvious form of reasoning, S may attribute to A beliefs that are slight variations of her own, for example as encoded in Rule PI2:

Rule PI2

$$BEL(S, EPLAN(A, \alpha_n, [\alpha_1, \ldots, \alpha_{n-1}], [\rho_1, \ldots, \rho_{n-1}], t_2, t_1), t_1) \wedge$$
$$BEL(S, CGEN(\alpha_n, \gamma, C_1 \wedge \ldots \wedge C_m), t_1)$$
$$\rightarrow$$
$$BEL(S, EPLAN(A, \gamma, [\alpha_1, \ldots, \alpha_n], [\rho_1, \ldots, \rho_n], t_2, t_1), t_1),$$
where $\rho_n = CGEN(\alpha_n, \gamma, C_1 \wedge \ldots \wedge C_{i-1} \wedge C_{i+1} \wedge \ldots \wedge C_m) \wedge$
$$HOLDS(C_1 \wedge \ldots \wedge C_{i-1} \wedge C_{i+1} \wedge \ldots \wedge C_m, t_2).$$

What Rule PI2 expresses is that S may ascribe to A a belief about a relation between act-types that is a slight variation of one she herself has. It asserts that, if there is some CGEN relation that S believes true, she may attribute to A a belief in a similar CGEN relation that is weaker, in that it is missing one of the required conditions. As another example, if S believes that two act-types α and β are quite similar, she thereby has reason to believe that it is plausible that A has confused them or has made a bad analogy from one to the other. Such reasoning is encoded in Rule PI3:

Rule PI3

$$BEL(S, EPLAN(A, \alpha_n, [\alpha_1, \ldots, \alpha_{n-1}], [\rho_1, \ldots, \rho_{n-1}], t_2, t_1), t_1) \wedge$$
$$BEL(S, SIMILAR(\alpha_n, \delta), t_1) \wedge$$
$$BEL(S, CGEN(\delta, \gamma, C), t_1)$$
$$\rightarrow$$
$$BEL(S, EPLAN(A, \gamma, [\alpha_1, \ldots, \alpha_n], [\rho_1, \ldots, \rho_n], t_2, t_1), t_1),$$
where $\rho_n = CGEN(\alpha_n, \gamma, C) \wedge HOLDS(C, t_2)$.

Rules PI1, PI2, and PI3 are merely meant to be suggestive of the sort of plan inference rules that can be stated within the analysis of plans as complex mental attitudes; a number of other such rules, along with examples of their use, can be found in chapter 6 of Pollack 1986. Such rules can enable a system to infer a wide range of plans, including plans that are constructed out of recipes-for-action that are not stored in the system's operator library. Adopting a view of plans as mental phenomena also makes it possible to reason about whether a particular set of actions is likely actually to be the object of an agent's plan. Consider once more the sample dialogue presented at the beginning of section 2. Imagine, however, that when the query is posed, S has reason to believe that A believes that Kathy is at home. In these circumstances S cannot use Rule PI1 to reason to the set of beliefs shown in figure 5.2; indeed, in this case A's query may be deemed to be incoherent. To distinguish between these two cases, it is necessary to have both a concern with the beliefs and intentions that are entailed by having a particular plan and a framework that distinguishes between the beliefs of the inferring agent and those that she attributes to the actor.

5 *Reconsidering the Traditional Model*

As I have emphasized throughout this paper, the traditional models in AI are models of the object of a plan, whereas my concern has been with the mental state of having a plan. But of course, when an agent has a plan, that plan does have an object—to wit, the set of acts he intends. What is the correspondence between the structure imputed to the actions that constitute a plan under the traditional approach and the structure imputed to the

actions that are intended by an agent that has a plan under my approach? More specifically, what correspondences, if any, are there between the relations between acts used in the traditional models—*causes*, *is-a-precondition-of*, and *is-a-way-to*—and those used in the account of plans I have developed here—*generation* and *enablement*?

The relations used in the traditional view involve some redundancy, as illustrated by the following two operators:

Header: flip switch	Header: turn on light
Effect: light on	Body: flip switch

These operators encode the same information, provided we equate the act of turning on the light with the act of achieving that the light is on. In general, whenever some act α *causes* some proposition P, it is also true that α *is-a-way-to* achieve(P). It will thus be sufficient to restrict our attention to the two relations *is-a-precondition-of* and *is-a-way-to*.

Consider then the following generic operator, in which α *is-a-way-to* β, and P *is-a-precondition-of* β:

Header: β
Preconditions: P
Body: α

Is there a way to map the relations expressed in it into the two relations *generation* and *enablement*? In fact, there are a number of plausible mappings. Figure 5.3 shows some of these.

Sentence 1 of figure 5.3 encodes the interpretation, apparently implicit in certain operators, that the preconditions P are sufficient for the performance of the header β by performance of the body α, but without necessarily being sufficient for the performance of α itself (and consequently, without necessarily being sufficient for the performance of β itself). This is the interpretation that most straightforwardly translates into the relations between act-types used in this paper: under this interpretation, the entire generic operator is associated with the sentence CGEN(α, β, C). However, it is not always the interpretation implicit in existing planning work. Sentences 2a and 2b encode the interpretation of the preconditions P as necessary for the occurrence of β: these two sentences differ from one another in the time interval during which P is meant to hold. Consider a typical planning operator in which a proposition representing that the power is on is included as a precondition for an operator with a header representing the act of turning on the light. It is not always obvious whether the former is intended to be a necessary, or merely a sufficient, condition for performance of the latter. (It might be merely sufficient, if, say, the light is attached to an emergency generator.) The preconditions in an action operator are sometimes also meant to be related to the performance of the act-type in the body of the operator—either to be necessary for it, as encoded in sentences 3a and 3b, or to be sufficient for it, as encoded in sentences 4a and 4b. One illustration of the preconditions being related to the body act-type α instead of the header act-type β would be an operator with a header representing turning on the light, body representing flipping the switch, and precondition representing both the power being on and the agent standing near the switch. The agent may well be able to turn on the light without standing near the switch—say, by throwing something at it. Standing near the switch is meant, in an example like this, to be necessary for the body action to occur. Combinations of the sentences shown in the figure are also possible: the interpretation underlying Sacerdoti's (1977) work seems to be a combination of sentences 1 and 4a.

Unfortunately, much of the existing planning literature has been vague about the intended interpretation of action operators and has used them at different times to mean different things. It is because of the resulting

Given

Header: β
Precondition: P
Body: α

Does this mean

1. $CGEN(\alpha,\beta,P)$, that is,
   $OCCURS(\alpha,G,t) \wedge HOLDS(P,t) \rightarrow OCCURS(\beta,G,t) \wedge \ldots$

2a. $OCCURS(\beta,G,t) \rightarrow HOLDS(P,t)$
2b. $OCCURS(\beta,G,t) \rightarrow \exists t_0[MEETS(t_0,t) \wedge HOLDS(P,t_0)]$

3a. $OCCURS(\alpha,G,t) \rightarrow HOLDS(P,t)$
3b. $OCCURS(\alpha,G,t) \rightarrow \exists t_0[MEETS(t_0,t) \wedge HOLDS(P,t_0)]$

4a. $HOLDS(P,t) \rightarrow \forall G[EXEC(\alpha,G,t)]$
4b. $HOLDS(P,t) \rightarrow \forall t_1[MEETS(t,t_1) \rightarrow EXEC(\alpha,G,t_1)]$
   where some further restriction on the length of t_1 is also given

5a. $HOLDS(P,t) \rightarrow \forall G[EXEC(\beta,G,t)]$
5b. $HOLDS(P,t) \rightarrow \forall t_1[MEETS(t,t_1) \rightarrow EXEC(\beta,G,t_1)]$
   where, again, some further restriction on the length of t_1 is also given

or some combination of these—say, Sentences 4a and 1?

Figure 5.3
Interpretation of the standard relations.

ambiguity of interpretation that I have avoided using the relations *causes*, *is-a-precondition-of*, and *effects* in this work and have instead made use of *generates* and *enables*. Any particular operator that makes use of the former set of relations can be expressed in terms of the primitive relations OCCURS and HOLDS—the two relations that were themselves used to define *generates* and *enables*. The translation must proceed on an operator-by-operator basis, however; because of the variety of ways in which the traditional relations have been used, there is no one translation that will apply uniformly to all the operators that have been proposed in the literature.

6 Conclusion

For nearly two decades, research on cooperative conversation has entailed a concern with the notion of plans and with the process by which one agent can infer the plans of the other agents with whom she is conversing. In this paper I have argued that in understanding plan inference, it is important to go beyond studying the structure of the object of an agent's plan: it is necessary to analyze the nature of the mental state of the plan itself. I presented such an analysis, in which having a plan amounts to having a particular configuration of beliefs and intentions. I then showed how this analysis can be put to use in a model of conversation that avoids certain limitations of any approach that does not consider the nature of plans as complex mental attitudes. In particular, I showed how an analysis of plans as complex mental attitudes makes it possible to reason about the plans of an actor even when those plans are invalid: how to reason about whether it is likely that the actor has an invalid plan, how to reason about the ways in which the plan may be invalid, and how to reason about what information to consider including in a cooperative response to a query that arises from an invalid plan.

When plans are viewed as complex mental attitudes, the process of plan inference can be seen as that of attributing a collection of beliefs and intentions to an actor. 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. I associated such discrepancies with a judgment that the actor's plan is invalid. Then I showed that the types of any invalidities judged to be present in a plan inferred to underlie a query affect the content of a cooperative response. I further suggested that, to guarantee a cooperative response, an inferring agent must attempt to ascribe to a 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. 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. Finally, I compared the representation used in traditional AI models of plans with the representation used here for the objects of an agent's plans, showing that the latter can be used to express and make more precise the intended meaning of uses of the former.

Notes

1. Recent work by Georgeff and Lansky (1986) is an exception.
2. For example, restrictions on the types of parameters, on the relations between parameters, and on the ordering of the subactions into which an action is decomposed. Constraints can be thought of as preconditions that the planning agent never attempts to achieve.
3. Several of the plan inference systems have additional rules that apply only to nodes that encode information-seeking actions. For example, many systems will construct a link from a node encoding "finding out whether P" to one encoding "achieving P." This construction is meant to capture the intuition that if an agent wants to know whether P is true, he may want P to be true. Kautz (1985) has shown how these rules, along with rules for handling nested plan inference, can be viewed as special cases of the three relations shown in figure 5.1.
4. This rule should be seen not as a sound rule of logical inference but as a rule that suggests an inference that is likely to be true. See Kautz and Allen 1986.
5. Condition 1, as well as Conditions 2 and 3, can be applied in "either direction": they can be used to add β when α is already in the subgraph, and they can be used to add α when β is already in the subgraph. Allen's Precondition-Action Rule, given above, corresponds to the former case for Condition 1. His system also includes the converse rule:

   SBAW(ACT) → SBAW(P), if P is a precondition of ACT.

6. Actually, there seems to be some tension regarding what is really meant by the W operator. Allen states that AW(P) means "A has a goal to achieve P," which seems to imply that P is a single action or property, not a whole plan. Consistent with this, he says that "SBAW(X) → SBAW(Y)" should be taken to mean that "if S believes A has a goal of X, then S may infer that A has a goal of Y" (1983, 120). But he uses these rules to infer not just that A has a goal of Y (that is, that his plan contains Y) but also that it contains X and Y related to one another in some particular way specified by the rule. So the rule relation preconditions and actions mentioned above should probably be written "SBAW(P) → SBAW(P → (is-a-precondition-of) ACT), if P is a precondition of ACT." Writing the rule this way would clarify Allen's model, but it would not affect the claim that the B and W operators are transparent to the inference rules.
7. Strictly speaking, the truth or falsity of the precondition at the time of the inference should not be what is at issue, but rather its truth or falsity at the time that the agent executes the action. Since the traditional systems have no explicit representation of time, however, this distinction collapses for them.
8. Throughout the rest of this paper I will make the simplifying assumption that when one dials a hospital, one reaches directly the person with whom one wants to speak. If the reader is uncomfortable with this simplification, the action of asking for Kathy can be inserted before the action of saying to Kathy "How are you doing?"
9. In fact, this condition may be slightly too strong: the agent need not be sure that performing his plan will entail performing his goal. In the normal state of affairs,

though, he will at least think this likely and will act as if he believed it. See the further discussion in section 3.3.

10. Bratman (1987) discusses the significance of partiality of plans in resource-bounded agents like humans and robots.
11. Notice that Π in this example need only be partially ordered: I may consider it irrelevant whether I first find out the phone number of St. Eligius and then the phone number of bank, or vice versa.
12. This claim may be less clear for the acts of asking Kathy how she is feeling and calling St. Eligius: it seems possible to say that I plan to ask Kathy how she is feeling by calling St. Eligius. The ordinary-language test is a rough one and occasionally fails to correlate with the phenomenal distinctions I want to draw.
13. Davidson's case rests upon examples such as the following: I might intend to make ten legible copies of what I am writing by pressing hard on carbon paper, without believing with any confidence that I will succeed. Grice maintains that many such examples are actually elliptical versions of conditional intentions. Contra Grice, Davidson argues that such an intention is *not* an elliptical version of a conditional intention to make the ten copies if I can, for since one cannot intend to do what is impossible, intending to do X if one can is equivalent to intending to do X simpliciter; nor is it an elliptical version of some more detailed conditional intention to do X if, for example, the carbon paper is particularly good, my hand muscles are more powerful than I thought, and so on. His argument against this is that "there can be no finite list of things we think might prevent us from doing what we intend, or of circumstances that might cause us to stay our hand" (1971, 94). This is obviously a description of the notorious "frame problem" that plagues AI.
14. Throughout, all variables should be taken to be universally quantified with widest possible scope, unless otherwise noted.
15. In fact, it captures more: to encode Clause 1 of Definition P0, the parameter i in Clause 1 of Defintion P1 need only vary between 1 and $n - 1$. However, given the following relationship between EXEC and GEN,

$$EXEC(\alpha, A, t) \wedge GEN(\alpha, \beta, A, t) \rightarrow EXEC(\beta, A, t),$$

the instance of Clause 1 of Definition P1 with $i = n$ is a consequence of the instance of Clause 1 with $i = n - 1$ and the instance of Clause 2 with $i = n - 1$. A similar argument can be made about Clause 3.
16. Recall the discussion in section 3.3 about assuming that an intention to α entails a belief that the agent will do α.
17. The plan inference rules should really be stated in terms of plausible eplans, that is, collections of beliefs and intentions that S thinks A plausibly has. When S has found some set of these that is large enough to account for A's query, their epistemic status can be upgraded to beliefs and intentions that S will, for the purposes of forming her response, consider A actually to have. See Pollack 1986, 126–130.
18. A rule symmetric to Rule PI1 is also needed since S can reason not only about what acts might be generated by an act that she already believes A plausibly intends but also about what acts might generate such an act.

References

Allen, James F. (1979) A plan-based approach to speech act recognition. Technical Report 121, Department of Computer Science, University of Toronto, Toronto, Ont.

Allen, James F. (1983). Recognizing intentions from natural language utterances. In Michael Brady and Robert C. Berwick, eds., *Computational models of discourse*. Cambridge, MA: MIT Press.

Allen, James F. (1984). Towards a general theory of action and time. *Artificial Intelligence* 23, 123–154.

Bratman, Michael E. (1983). Taking plans seriously. *Social Theory and Practice* 9, 271–287.

Bratman, Michael E. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Brown, John Seely, and Richard R. Burton (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science* 2, 155–192.

Carberry, M. Sandra (1985). Pragmatic modeling in information system interfaces. Doctoral dissertation and Technical Report 86-07(1986), University of Delaware, Newark, DE.

Collins, Allan, Albert Stevens, and Sarah E. Goldin (1979). Misconceptions in student's understanding. *International Journal of Man-Machine Studies* 11, 145–146.

Davidson, Donald (1980). Intending. In *Essays on actions and events*. New York: Oxford University Press.

Fikes, R. E., and Nils J. Nilsson (1971). STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence* 2, 189–208.

Genesereth, Michael R. (1979). The role of plans in automated consultation. In *Proceedings of the Sixth International Joint Conference on Artificial Intelligence*, Tokyo.

Georgeff, Michael P., and Amy L. Lansky (1986). Procedural knowledge. *Proceedings of the IEEE, special issue on knowledge representation*, 1383–1398.

Goldman, Alvin I. (1970). *A theory of human action*. Princeton, NJ: Princeton University Press.

Grice, H. P. (1971). Intention and uncertainty. *Proceedings of the British Academy* 57, 263–279.

Kautz, Henry A. (1985). Toward a theory of plan recognition. Technical Report 162, University of Rochester, Rochester, NY.

Kautz, Henry A., and James F. Allen (1986). Generalized plan recognition. In *Proceedings of the National Conference*, American Association for Artificial Intelligence, Philadelphia, PA.

Litman, Diane (1985). Plan recognition and discourse analysis: An integrated approach for understanding dialogues. Doctoral dissertation and Technical Report 170, University of Rochester, Rochester, NY.

Nilsson, Nils J. (1980). *Principles of artificial intelligence*. Palo Alto, CA: Tioga Publishing Co.

Pollack, Martha E. (1986). *Inferring domain plans in question-answering*. Technical Report 403, SRI International, Menlo Park, CA. Also Doctoral dissertation, University of Pennsylvania, Philadelphia, PA.

Sacerdoti, Earl D. (1977). *A structure for plans and behavior*. New York: American Elsevier.

Sidner, Candace L. (1985). Plan parsing for intended response recognition in discourse. *Computational Intelligence* 1, 1–10.

Woolf, Beverly, and David McDonald (1983). Human-computer discourse in the design of a Pascal tutor. In *Proceedings of the Conference on Human Factors in Computing Systems*, Association for Computing Machinery's Special Interest Group on Computer and Human Interaction, Boston.