Discourse as planned action

Samuel William Dyne Steel

PhD

University of Edinburgh

1984

Abstract

A remark fits into a discourse if it can be interpreted as a plan-advancing act given the world around the speakers, their intentions, and the prior discourse. Teleological questions about discourse such as "Why did X make this utterance here?" must have answers such as "Because X wanted to do so-and-so". When there is no such answer, the discourse will be incomprehensible.

The proper sort of explanation to give is to talk about the speaker's plans and the alterations he intends in his interlocutor's plans. To do this one must have an account of what a plan is, how it can be changed by changes in its maker's beliefs and values, how utterances can make such changes, and when such changes count as benefits for the speaker. Such an account is offered.

The interlocutors must have recursive beliefs (beliefs about each others' beliefs). But often a special simple case of recursive belief can be used. Then recursive belief can be factored out of the problem of explaining how utterances change plans. A proof of sufficient condition for this special case to arise is given.

Indirect communication can occur if a speaker forces his hearer to change the plans the hearer supposes the speaker to have. This happens because being known to have a plan entails being known to have certain beliefs.

Some discourse events are constituted by changes that occur to the plans of speakers in the discourse as a result of what is said. Examples are given.

Some process accounts of the recognition of utterances as goal-directed attempts to change plans are considered.

I want to thank my wife, Louisa, and my supervisor, Henry Thompson, without whose different assistances this thesis would never have been written.

I declare that the work in this thesis is entirely my own.

CONTENTS

Introduction

The story of this thesis is that if you look at language use as purposive action, you have an interesting way of explaining meaning and discourse. If you take a discourse, and point to some utterance in it, you can ask "Why is this remark here? What does it mean here? Why this one and not another one? Why any remark at all?". I suggest that this is the same as asking "Why did the speaker say this here rather than that? What did he mean by saying this here? Why did he say anything at all?" and that the proper sort of explanation to give is to talk about the speaker's intentions in terms of his plans and their intended effect on his interlocutor's plans.

I suggest that a remark fits into a discourse if it can be interpreted as a rational plan-advancing act given the world around the speakers, their intentions, and the prior discourse. Teleological questions about discourse such as "Why did X make this utterance here?" must have answers such as "Because X wanted to do so-and-so". When there is no such answer, the discourse will be incomprehensible.

The structure of the thesis is this:

Chapter 1 states what I am trying to argue, which is that discourse must be seen as planned action. It lists some aspects of discourse that seem inexplicable without doing this: for instance, some exchanges in which the parties must have inferred each others' plans; or the occurrence of events in discourse which are not definable in terms of the literal content of what was said. It argues that it is impossible to imagine methods of communication that could be used by independent agents which are not based on seeing utterances as actions. Any other method seems to have to rely on speakers being

able to force beliefs and values (albeit very abstract ones) on their listeners. and would apparently let speakers make listeners into mere tools of their interlocutors. Lastly it briefly discusses the relation of the function-type of sentences and the use to which they are put.

Chapter 2 is on related work. It considers the work of speech act theorists such as Austin. Searle and Grice. Much of this is accepted. but not their emphasis on the importance of illocution. Then it covers the process accounts of Allen. and contrasts what I propose with what he offers. Lastly it denies the possibility of giving any deep explanation of discourse by mapping it onto scripts of the sort that Schank proposes.

Chapter 3 gives a model of rational action. It describes actions in a way that lets them be used in proofs in the same way as sentences describing states of affairs can be. It argues that rational action can be seen as action that is part of a plan to a desired goal. and that a plan can (for my purposes) be equated with a particular sort of proof. Finally it briefly considers and rejects a representation of plans that emphasizes proving the desirability of means (actions). over proving the possibility of ends.

Chapter 4 links action and utterance. It argues that utterances can force a rational planner to remake his plans in the light of the changes in his beliefs and values that the utterances provoke. It shows how plans seen as proofs are vulnerable to new information. Lastly it shows when changes in intended plan are good for the planner, even when. as sometimes happens. the information destroys the possibility of a good plan being devised.

Chapter 5 is about indirect communication. It argues that knowing

that a planner has a plan entails certain things about the planner's beliefs and values. Changes in what someone takes his plans therefore entail changes in what that someone must take his beliefs and values must be. So a planner can change what is known about his plans as a way of compelling others to change their beliefs about him. The planner can use these changes to do exactly the same thing as he might with overt statements of what he wants and believes.

Chapter 6 discusses belief about other peoples' beliefs. Clearly this is vital to indirect communication, and indeed to practically everything else in discourse. Fully general recursive belief, where for instance A thinks F while A thinks B thinks A thinks -F is in fact necessary. But in a large range of situations, a simpler special case of recursive belief, mutual belief, can be used instead. When this is possible, the recursiveness of recursive belief can essentially be ignored. A proof is given of a sufficient condition for this. It is shown that much discourse fulfills this condition, and so plan manipulation can be discussed without reference to recursive belief.

Chapter 7 considers series of utterances that affect plans, and suggests that certain speech acts and discourse events are constituted by series of certain sorts of changes in interlocutors' plans. Therefore recognition of these is only possible if the series of changes made to a plan can be recognized. Two sets of examples are given: offers, requests and responses to these; and explanations, which are construed as changes of plans such that the explicandum is included in the changed plan in such a way that it is a reasonable event or state to have expected.

Chapter 8 is a sketch of two processes that could detect the changes being attempted by the maker of an utterance. One uses a fixed case

analysis of all the changes that could be made by an utterance, and suggests that understanding an utterance is finding a path through the analysis compatible with the utterance and the currently extant plans. The other uses chains of inferences built from rules whose antecedents are purposely not checked. If the chains that can be built starting from the utterance and an extant plan run to a common "interesting" fact, it may be wise to see if these chains can be made into proofs. If they can, a wise planner will make certain changes to his plans. These changes may be credible purposes of the utterance. These approaches are both rejected, but only as means of realizing the theory, not in a way that impugns the theory itself.

Chapter 1. Some problems in discourse

Let me ask some questions about language, and then give some answers that make me think one needs to think about the speaker's intentions. Be warned; this is an explanation of why I think intentions matter. I am not going to give processes for answering all these questions if they are asked about specific text.

1.1. Some questions

1.1.1. What is it that matters in what is said?

Even for simple statements of fact that seem only to have the purpose of giving information, the important information that one gets from them is not the information borne in their truth conditions. Here is an example:

> A and B always make tea in a teapot with water boiled in a kettle, and always make coffee in a percolator.
>
> A: Do you want coffee or tea?
>
> B: I've put the kettle on.

B means all sorts of things by his reply. One of the things he means is that he wants tea. This is not of course any part of the truth conditions of what he said. Is it fair to say that the knowledge that B wants tea is part of the meaning of what he said? I think so. Imagine asking A whether B wants tea or coffee after this exchange. A will say "tea". A moment before he didn't know that, or he wouldn't have asked his question. The only place his information can have come from is from what B said. So it is a piece of information A had to

know before he could be said to have understood what B said, and should therefore count as part of its meaning.

So more ingredients than just truth conditions are involved in fixing the meaning of an utterance. What are they?

And of course questions and commands don't have truth conditions. Even though they may involve pictures of the world, they don't present it as a picture of how the world is. Questions and commands are intended to provoke some sort of response from the hearer. Statements are supposed to provoke belief; questions, a reply; commands, obedience - at least in some sort of ideally simple discourse. The remark is supposed to tell you its intended response by its mood. But of course overt mood doesn't really tell you that. "Can you shut the door?" is often a (polite) command posing as a question. In

A: Where is the key?

B: Isn't it on the mantelpiece?

B's remark is a (hesitant) statement posing as a question.

One hasn't understood what is said until one has found the intended response. How is it created by a speaker or detected by a hearer?

1.1.2. What use is language?

Why should we talk at all? We have got the ability, so presumably it does something for us as a species. What? One answer might be that it gives us remote perception. Other people see things, and then they tell us that they have seen them. I could now tell a story about how useful this sort of warning of predators and advice about food would

be to a tribe of apes.

The story could be true, but is it relevant? Suppose language were primarily for finding food and avoiding tigers. All one would need would be a table saying "If you see honey then make the honey call, if you see tiger make the tiger alarm call". This would achieve the species' simpler aims without all the effort involved in being able to describe arbitrary situations. But if the species does (to speak loosely) decide to go for general descriptive ability, so that it can talk about anything it perceives, this approach wouldn't work. A speaker will have an effectively infinite number of things that he can say about any situation.

The ape needs not only a code for describing whatever it is that its hearer can't see, but also a method of deciding what parts of what it can see to relay to its hearer. What method?

1.1.3. What makes discourse cohere?

The order of remarks in a conversation matters. One can't scramble them and leave their value the same, as one can with a pocketful of change. This discourse doesn't make sense:

B: Isn't it shut?

A: Waverley.

B: I thought it was a holiday today.

A: The strike's over.

B: Let's go then.

A: Not for the trains.

B: Which one?

A: Can you give me a lift to the station?

even though it just a scrambling of:

A: Can you give me a lift to the station?

B: Which one?

A: Waverley.

B: Isn't it shut?

A: The strike's over.

B: I thought it was a holiday today.

A: Not for the trains.

B: Let's go then.

What is it that makes one acceptable and the other incoherent? What sort of rules tell you that one remark will fit into the next turn in a conversation while some other won't? What leads one to say something rather than stay silent?

1.1.4. How do we define and apply words that describe discourse?

There are all sorts of words that we use to describe events that occur in discourse: rebuke, threaten, promise, convince. These are actions; we speak of people doing them. What sort of definitions should these words have? And once they have definitions, what part of the world should one look at to see whether the definitions apply?

1.1.5. Why can non-linguistic actions behave like utterances?

Suppose I cut my finger at some moment when I can't speak. (Perhaps a tape-recording is being made.) You know where the sticking plaster is. I wave my bleeding finger at you. You go and get the plaster and give me a bit.

In that scene I have got you to do exactly what you would have done

if I had told you I had cut myself. In some sense I have done the same thing as if I had asked you outright for the plaster. But I didn't use language. What is it that linguistic and non-linguistic actions have in common that lets both affect other peoples' actions?

1.2. Some answers

1.2.1. What is it that matters in what is said?

The example was:

> A and B always make tea in a teapot with water boiled in a kettle, and always make coffee in a percolator.
>
> A: Do you want coffee or tea?
>
> B: I've put the kettle on.

What B means is "I want tea". The knowledge that B prefers tea is deducible from A and B's shared assumptions about how they make tea and coffee and B's remark, taken with some axioms about rational behaviour. (Expanding that clause is the difficulty). One can argue that a person only does intentionally those things that he thinks will benefit him. Putting on the kettle will only benefit B if he wants tea. So one can deduce B wants tea.

That is indeed the intuitive meaning of B's remark, and it is a deduction from what B said. But why is that deduction the one that matters? Should one say that the meaning of a remark is all that can be deduced from it? Surely not. For instance, from "F" one can deduce "F or G". But "B prefers tea or the moon is made of green cheese" is no part of the meaning of what B said.

"B wants tea" is the deduction that matters because it is the one that can be seen to have a purpose: it answers A's question.

(There is a possibility of confusion here. I am not saying that this deduction is the right one because it TALKS ABOUT B's intentions (to make tea). It is the right one because it ACHIEVES one of B's intentions (to answer A's question). An example such as

A: Is Dumfries in England or Scotland?

B: It has a team in the Scottish football league.

has exactly the same structure as the previous one, but makes no reference to anyone's intentions. The added information allows A to deduce an answer to his question.)

What does this have to say about meaning? If the story I am telling is true, then perhaps meaning adheres to utterances, not sentences (not a novel claim), and perhaps meaning can be identified with the change of belief that the speaker intends to take place.

But there are of course utterances that aren't intended to change the hearer's beliefs. The standard sort of question, where the speaker hopes only for the answer and nothing else, or the standard command, where he hopes for obedience and nothing else, are like this. Do questions and commands then not have meaning, or have a radically different sort to statements' meaning?

I suggest that the central task of extracting what we feel to be the content of an utterance is not finding out how it changes our beliefs, or indeed even how the speaker intended it to change our beliefs, but finding out what effect the speaker intended his utterance to have. What he intended may involve the hearer changing

his beliefs. It may be sensible to identify utterance meaning with the set of all the changes in belief that utterance causes. But this set of changes is the means, not the end, of the utterance.

Of course in many cases the intended effect of an utterance is the change of belief it causes. If I am teaching someone about spherical harmonics what I care about is the new beliefs that he entertains about them. What he then does as a result of this information may well be of not the slightest interest to me. Even when I make some statement because of the effect I hope it will have on my hearer's actions, I will have a subsidiary intention that he believe that statement. Without that change of belief there will be none of the other, more important, effects.

(To be a little polemical, perhaps this is why naive semantics sometimes concentrates too much on trying to relate sentences to the pictures of the world that they express. Often the change that happens if the hearer believes what he is told IS the change that the speaker intended. If you make the mistake of thinking that they are always the same, and that therefore bending the hearer's picture of the world so that it is the same as the picture expressed in a sentence is the central task of language, you will think that truth conditions are meaning.)

Questions and commands clearly are used by speakers for their effects. Nevertheless, one can ask a question such as "What did Smith mean when he asked "Are there any matches left?"?" or "... when he said "Bring me the matches!"?". If the answerer can overcome his urge to say "What do you mean "What did he mean?"?", you may well get an answer such as "He meant he wanted to know if there were any matches left" or "He meant he wanted me to bring him the matches". I can explain this by saying that if the hearer has explained Smith's

question by assuming that Smith hopes to get an answer or the matches, the hearer must believe that Smith wants an answer or the matches. These are new beliefs to the hearer, and so part of the meaning of Smith's demands.

There are of course already accounts of how utterances are like acts. for instance from Searle and Austin. These are discussed in chapter 2, and I give reasons there against adopting them. Briefly, it is because they adopt a special kind of action, illocutionary acts, as explanatory of the effect that utterances have on their hearers. Then the issue is how utterances can realize illocutions and how language users can detect such actions in discourse. I would rather drop illocution in favour of claiming that utterances have their force because of the effect that their propositional content has on mutually known plans. Such effect should be predictable just from considering the nature of plans and how they should be rationally altered when the context in which they were made alters.

The existing approach most similar to this is that of Ellman (1983), who proposes to deprive utterances of all special force except that of making their content mutually known. After this, an agent will infer the indirect goals of the speaker using rules that are just about planning and which make no reference to speech act force.

> "As hearers may choose not to accede to REQUESTs, or decide to cooperate in other ways, it would seem that REQUESTS only suggest, or INFORM, the hearer as to the speaker's desires. [...] That is, from the point of view of an understander of speech acts it may not be relevant that utterances are REQUESTs, as long as the hearer can detect from them the speaker's superordinate goal(s). It therefore seems that the notion of different speech act types is not necessary for

natural language understanding." (Ellman 1983).

I think he and I would differ in principle only on the primitive status of mutual belief, and the machinery by which speakers' intentions are recognized.

1.2.2. What use is language?

Things are useful to us if they achieve our goals or prevent our fears. If you think of utterances as actions capable of affecting the world, albeit just that small part of the world that is other people, then one can explain the fact that we say some things and not others by saying that we say only those things that advance our ends, because we want to change either others' belief, or others' actions based on their beliefs.

The ape's problem about what it should say in given circumstances is solved if it can guess both what its fellows are likely to want and to fear, and what pieces of information are likely to matter to them because of their wants and fears. It can then restrict itself to offering only information that it knows its fellows will be glad to have because it helps them obtain their ends, or which will make them help him.

1.2.3. What makes discourse cohere?

The answer to this is related to the last. To repeat what I said in the introduction, I suggest that a remark fits into a discourse if it can be interpreted as a rational plan-advancing act given the world around the speakers, their intentions, and the prior discourse. Teleological questions about discourse such as "Why did X make this utterance here?" must have answers such as "Because X wanted to do

so-and-so". When there is no such answer, the discourse will be incomprehensible.

(Let me put in a disclaimer. There are lots of things one can say about discourse. I am not claiming that ALL discussions of it have to be teleological.)

1.2.4. How do we apply words that describe discourse?

It seems that at least some of the words that we use to describe what happened in discourse are defined in terms of plans to change people's beliefs and actions via what has been said. For instance, consider the word "rebuke". Suppose B parks his car in front of a fire exit, with possible evil consequences. What has to happen for B to have been rebuked by A?

We can't establish that a rebuke has occurred just by looking at B at two moments, and finding that in the first B approved his action, but later didn't. B might have parked his car, and later have discovered from C that it was in front of a fire exit.

Nor by finding a chain of events such that an action of A's led to B disapproving of his own action. B could have parked his new car that A has never yet seen, and then hear A say "Look at the car some fool has parked in front of a fire exit", and so come to disapprove his own action.

Nor by finding a chain of events such that an action of A's that A intended to lead to B disapproving of his own action succeeded. A could have advised a fire officer to remonstrate with B, with the intended effect. This would not

be a rebuke by A.

One has to see a series of actions as the execution of a plan of A's, by which A says something to B in order to change B's beliefs, so that B come to believe that B's action X was an action with foreseeable bad effects. For instance, if A says "If there's a fire they won't be able to get out of that door", that remark then counts as a rebuke.

Notice that A's success is irrelevant. If B replies "Tough" to the rebuke, it is still a rebuke.

Why does knowing how to apply such terms matter? Since we are social beings, to whom being approved is almost as important, though not as tangible, a goal as being fed, we will act so as to end up approved. I may forego my own convenience to give your friend a lift, in order to get your goodwill. This is an example of us acting in a world that contains socially as well as physically real objects. But socially real states of affairs are emergent on what people do and say and believe, not events in nature like its starting to rain. We can alter social reality by what we say - as for instance in rebuking people. Such social states may condition our future actions, both within and outside discourse. If I know I have been rebuked, I may attempt to defend myself; I may apologize. We must be able to spot such events to either participate in or understand later discourse. But (as with rebuking) we can only spot these events if we see actions and utterances as parts of plans.

1.2.5. Why can non-linguistic actions behave like utterance?

I suggest that an action is an attempt at communication if it is part of the execution of a plan intended to influence your actions by

trying to change your belief and values. If this is so, the means employed in the plan are irrelevant. Changes in your belief brought about by showing you bits of the world (say a bleeding finger), and by telling you that that is how the world is, will be indistinguishable.

1.3. The effect on communication of looking for speaker's purpose

Given that utterances are going to be explained as goal directed, how will language accommodate this? To explain an act or an utterance as rational means to see it as part of the execution of a plan. To suppose that a person is following a plan involves supposing that he holds the beliefs that would make such a plan likely to succeed, and the values that would make it worthwhile.

But explaining an utterance isn't just a matter of seeing an utterance as part of a plan already known about in detail. One may have to ascribe new plans to a speaker. If these new plans are to be explanatory, one has to stick with the new commitments about the agent's beliefs and values that they force on you. For instance, if I explain A's question "Are the banks open?" as part of A's plan to cash a cheque, I must suppose that A wants to cash a cheque, believes he has an account and so on. The change in my beliefs and values that occurs when I accept that A has this plan is exactly the same as if A had told me about them explicitly.

Now a cunning speaker may take the opportunities this this offers him. He may make remarks whose purpose is not the change that the sentence might be expected to make, but the changes that follow from the assumptions that its explanation require. As a result, no understanding directed to seeing what changes the speaker intended to make in you can be complete until you've looked for changes of this

sort too. But of course the speaker will not always be doing this. One can't tell prima facie whether he is or not, so one has to look for the utterance's purpose every time.

1.4. Utterances are like actions

These answers are by no means complete answers. But I believe they show an interesting common element: the importance of seeing utterances as rational action. Utterances are like actions in these ways:

> Actions change the world. There are some sorts of changes, changes in another person's mind, that utterances are particularly useful for bringing about.
>
> Actions change the world. Usually you have to make changes in the physical world yourself. But sometimes you can make changes in another person's mind so that he will act so as to make the changes that you want in the physical world.

Further, we are capable of understanding events and actions. We endlessly try to tell ourselves stories that make our experience cohere. Any machinery we deploy to do this will be equally available to explain utterances.

1.5. What is the effect of utterances?

In this section I want to discuss, first, why it seems odd that language should have any effect on people at all; and secondly, some reasons why, despite this oddness, it does have an effect. What follows is inconclusive; I have not got the answer; but the problem and some attacks that can be made on it are I think worth describing.

Briefly, the difficulty is that any plan recognition process has to have as input some facts about what the doer of an action expects its effects to be; but when the action is an utterance, it is hard (for reasons to be given soon) to see ANY sort of effect that a rational agent could expect his action to have.

1) When I push too hard on a pane of glass, it breaks because of the force I apply. The glass has no say in the matter. Are utterances like this? Do they have an irresistible effect on the person who hears them, at any rate beyond the purely physical? It is hard to see how this can be so without saying that we are compelled to believe what we are told.

2) It is hard to see how there can be a systematic pairing between an utterance and the particular change in beliefs and values it might be intended to produce (the utterance's content).

I try to suggest how utterances may be seen as compelling us to believe, and explained as if that was their purpose, without them actually compelling us.

1.5.1. Why a story involving compulsion is wrong

What is the alternative to utterances being understood via plan recognition? Restrict the discussion to languages in which all utterances are statements of fact about how the world happens to be at some moment. Would one then need to consider a speaker's intentions? Why can't one be content with the notion that utterances have literal meanings independent of their maker's intention, which

enable a language to do all that we want it to do? Here is a parable a person who asked such a question might tell:

> Imagine intelligent rational creatures; that is, creatures who perform actions which they believe will collectively transform the world as they think it is into the world as they want it to be. Imagine further that they see each other like this, and can envisage non-actual courses of action. Then they will be able to foresee how others' actions will depend on their beliefs. Now if they can cause each other to know about parts of the world they can't themselves see, they will have a partial control over the others' actions. The advantages of this for the controller (altruistic or selfish) are obvious. How can they do such "causing to know"? Well, they could if they had a system of simple action-types (call them "sentences") which they could perform at will, and which were systematically related to states of affairs, so that given a performance of one of these actions (call it an "utterance") they could find the related state of affairs and vice versa. Let's call this their "language". Now creatures like this "understand" an utterance made by someone else when they have found the state of affairs that it's associated with. Thereafter their actions may be based on the new states of affairs they believe in, but that's part of their rationality, not their language use.
>
> This (my opponent says) is of course a picture of human language. Where did I have to introduce recognition of speaker's intention?

I don't believe this is a picture of language. Furthermore, I don't

think it could be. The two objections to this sort of account are:

> What it is that a hearer does when he reacts to an utterance is obscure. He certainly doesn't just bend to what he's told.

> It isn't possible to have a pairing of sentences and states of affairs in a way that lets the account go through.

1.5.1.1. Problems with hearer's reaction

What does a hearer do with an utterance when he hears it? Here are two wrong answers.

1.5.1.1.1. The hearer does not have to believe

The hearer recovers from it a picture-in-the-head of a state of affairs. Next he believes it; he alters his beliefs about the world until they include the picture he has been presented with.

But this seems to make the detection of deceit impossible. Suppose A attempts to serve his own ends by getting B to injure himself by telling him one of B's intended actions was safe when it wasn't - eg that some vegetable was edible when it wasn't. How could B protect himself? He couldn't. A language that worked like this would kill its hearers.

1.5.1.1.2. The hearer is not able to check

Is it possible to evade the problem by supposing that the hearer applies some check to what he's told before he believes it.

He could compare what he's told with his own knowledge of the world. But if he's regularly in a position to do this authoritatively, there's no point in his ever listening to anyone, since he knows it all already; and if he doesn't, his check is powerless.

He may have a policy of believing only his friends, those who presumably always tell him the truth. If he can make such a distinction, he's all right. But surely his companions will sometimes lie and sometimes tell the truth. Nor could he tell whether a stranger should be listened to or not.

Similar arguments apply to the detection of error rather than deceit in what he's told - in the latter case, with the replacement of "friends" by "people who know". We need a different account of how utterances affect hearers.

1.5.2. Problems with pairing sentences and states of affairs

No state of the world can be described even as accurately as we can perceive it, or as our hearer can imagine it. The best one can do is offer a description whose range of applicability covers what we were trying to describe. There is no one picture that corresponds to a sentence.

But my opponent objects: "This is silly. Of course when I abstract my kitchen by saying "The sink is by the stove" I'm not attempting to describe my whole kitchen, or even the sink and the stove. All I'm doing is specifying a condition that separates pictures where the sink is by the stove from pictures in which it isn't. Your hearer isn't going to be able to recover the (so to speak) finished oil-painting you have in your head, buts he's got a pencil sketch of some of its features."

Here I want to follow Searle, who argues that this account fails too. "[T]he received opinion [...] errs in presenting the notion of the literal meaning of a sentence as a context free notion." (1978:210). He takes the sentence "The cat is on the mat", and then imagines a series of peculiar cases (cats on stiffened mats at strange angles, cats on both sides of mats in free fall) where we would be stumped to say whether "The cat was on the mat" was true or not. Then he goes on to consider cases where we are nevertheless able to use "The cat is on the mat" perfectly effectively in these funny situations; for instance, when we assert that the cat is not on the mat any longer if it had been closely apposed to the mat in free fall but is now two yards away. From this he concludes that there is NO fixed picture or sketch associated with "The cat is on the mat", and that our successful use of it must be explained another way.

He asserts that he is not denying that an utterance has a literal meaning, just that it cannot be recovered without consideration of the context in which it was made. "Literal meaning of a sentence only has application relative to the coordinate system of our background assumptions." (1975:220). He is in opposition to the accounts such as that of Fish (1979), where literal and non-literal meaning are equated. Fish claims that our recognition of the picture of the world that an utterance represents, and of the ulterior intentions of its speaker, are the same sort of thing, at different points on a cline. In theory-laden terms, the literal meaning of an utterance and the indirect speech act it realizes are the same sort of thing. The account claims that we exert the same faculty to understand that (in the proper circumstances) "The curtains are yellow" means that the curtains are yellow, and that "Would you like a cup of coffee?" is an invitation to sexual intercourse.

Where these accounts are alike is in claiming that utterances will not have a known effect which can be the uninterpreted data to feed into a plan recognition mechanism. In both, the claim goes, in order to recognize the content (whatever that is), one needs the vast knowledge we have as members of our culture. The amount of input beyond what we hear is as great whether we are deploying it to recover speech acts, speaker's intentions, or pictures of the world. Clearly one cannot see the recovery of literal meaning as a more or less determinate sub-task of the task of recovering content. Very likely literal meaning and speaker's intention or speech act are the same sort of thing, and the same process, sensitive to all that we know, recovers them both.

Such an argument seems to have one large flaw.

Either sentences have some minimal meaning independent of their use or they don't.

If they don't, then the process of interpretation will be insensitive to the utterance it is set to work on. All sentences will be equivalent. The output of the interpretation will depend entirely on our cultural knowledge. At any time any utterance will be as good as any other. While there are occasions that approximate to this, (eg when buying an underground ticket at rush hour I will quite likely construe anything the clerk says as "where?"), this isn't generally true.

If they do, then the only mechanism that I can suggest for the process, whose details have not yet been specified, is that it is a plan-recognition process based on the effects caused by recognition of this meaning: which is what I was

arguing initially.

Imagine a language where sentences had no meaning, however minimal, independent of their use. Another parable:

> Suppose the creatures, puzzled by the failure of their first communication system, try again. They say, let us build a picture from the remarks that our fellows make to us which is the one that the speaker intends us to make. Let us do this by asking ourselves, whenever we hear a remark, what he presumably did mean and respond only to that. We can do this (they say) because we can tell what he would say in that context if he was rational. Let us be confused by nothing else, in particular not by the supposed literal meaning of the sentences which in any case may be different from those the speaker intended them to bear.

What happens? They are no better off than before. One of them makes a remark to another as they go off shopping for a bath plug in Woolworths. "Woolworths has them" he says. The other tries to understand this. What, he asks himself, does my companion intend me to understand? We are engaged in a plan we both know about. We expect it to succeed if we carry on. There is no point in his speaking unless to tell me something important I don't know. What could that be? Aha! That the plan won't work, because Woolworths has none. That must be what he said - and so he abandons the enterprise.

So I have to reject both accounts like Fish's and like Searle's. Both prevent a plan recognition process starting up. I also have to reject an account like Fish's for another reason. In the account of plan interaction given later in this thesis, utterances are said to affirm or contradict nodes in a plan. Affirmation or contradiction can only

occur between propositions. So utterances must have propositional content. Searle agrees utterances have such content, but denies it is easily recovered. Fish denies they have such content.

1.6. Can plan recognition be applied to utterances?

At this stage, understanding utterances may look impossible. They can't work as if they were updates to a data base. They can't have an effect based purely on what happens to a hearer if he applies plan recognition. And these facts may seem to prevent plan recognition being applicable to utterances at all. The problem is this:

- The process of plan recognition takes an action, considers the effects, immediate or remote, that its doer may have expected it to have, and hypothesizes that one (or more) of these effects is the goal of the action.

- So plan recognition applied to an utterance must look for the effects that the utterer may have expected, and guess which is the goal of the utterance.

- The sort of effect that an utterance has (at the level of description that matters here) is to affect an agent's beliefs and values, so that the plans that he makes founded on those beliefs and values change too.

- Which change in beliefs and values occurs must depend on the propositional content of the utterance. It was argued earlier that a system in which the propositional content made no difference to the effect of the utterance would be impossible.

- But it was also argued that utterances had no irresistible effect on their hearer's beliefs and values, so the speaker can have no certain expectations about the effect of an utterance. The effect on the hearer is not dictated by the content of the utterance.

- So since plan recognition supposes that an agent has expactations about the effects of his actions, plan recognition is inapplicable to utterances.

I must reject the conclusion. How?

1.7. Plan recognition and actions that must fail

But there are non-utterance actions that look as if they depend on plan recognition but which don't fit into the pattern of plan recognition sketched above. Here is an example from my own breakfast table:

> A and B are sitting at breakfast. A has a cup of black tea in front of him. He usually has it white. The milk is not on the table, but on a shelf about two feet behind B, and about four feet from A. A stretches out his arm towards the milk, but clearly can't reach. B, seeing him stretching, reaches for the milk and passes it to him.

B guessed A's want, but on what grounds? It cannot have been by guessing that what A was trying to do was to reach the milk, because B must have known that A could have seen that he wasn't going to succeed, and so B can't think both that the stretching was part of a plan and that A is rational.

It seems we have a way of construing actions that must fail as if they were actions that might succeed, and then responding to the goals they will not in fact achieve. Such a way of construing actions may be applicable to utterances as well.

1.7.1. How abortive actions may seem to have effects

There are three possible accounts I can give about how actions that are necessarily abortive are responded to as if they might have succeeded.

1.7.1.1. Knowledge of conventional effects of actions

We all "know" that King Alfred burnt the cakes, and that Washington cut down a cherry tree. Whether these facts are true or not is irrelevant. We pretend, when in the right circumstances, for instance when being facetious about burnt cakes or felled cherry trees with people of our own background, that they are true. We know that these pretended facts are mutually known. We do not have to check that the other chap knows them; they can be relied on as part of our common mental furniture.

Perhaps we are able to treat actions as if their doers expected them to have some effect, even when the action, were it to be performed with no-one about to respond to it, would fail. In the same way as we know the story about the cherry tree, we all know the story that tells us that when someone makes an assertion to us, we alter our beliefs to allow it in. Whenever we say something, we expect our hearer to pretend that he believes that the assertion will make him change his belief, and then go on and guess what else we might mean and intend in the light of this pretence.

For instance, if B thinks that he is expected to treat A's stretching for the milk as if he thought that A expected he would reach the milk, then B can guess what effect A expected his action would have, and B can apply plan recognition to grasp, and perhaps assist, A's plan.

(If instead what one did in such cases was to treat actions as if they WERE going to have some effect, not just as if their doer expected it to have that effect, plan recognition would not help recover agent's intentions. If B treated A's stretching for the milk as if A was going to reach it, there would be no need for B to assist A, and A would end up without the milk.)

Now straight plan recognition works because the recognizer has knowledge that lets him associate seeing an act with the effects the doer expects the act to have. The sort of pretending plan recognition described here will only work if the recognizer has knowledge that lets him associate the act with the effects that he should pretend that the doer thinks the act will have.

Such knowledge could be called knowledge of the conventional effects of actions. We have conventional knowledge of WHEN an action should be performed regardless whether we perform it at such times: we know that the proper time to change into a dinner jacket is just before dinner, even if we never do so. Similarly, we have conventional knowledge about the EFFECTS of actions - for instance, about what happens if one tells an improper story in the presence of a clergyman, whether or not such effects actually occur. Jokes can be based on such knowledge.

It is not the acts referred to that are conventional, it is our apparent (but of course non-existent) agreement to act as if those

actions had those properties. Such an apparent agreement is close to (but not the same as) the definition of convention offered by Lewis (1969:78) of which the relevant parts are

> A regularity R in the behaviour of members of a population P when they are agents in a recurrent situation S is a convention iff it is true that, and it is common knowledge in P that, in almost any instance of S among the members of P
>
> 1) almost everyone conforms to R;
>
> 2) almost everyone expects almost everyone to conform to R;
>
> [...]

It is this closeness that suggests calling such knowledge "conventional".

Lewis also argues that the actual falsity of the beliefs that it is conventional to act as if we believed doesn't matter.

> "A fictive precedent would be as effective as an actual one in suggesting a course of action for us, and therefore as good a source of concordant mutual expectations enabling us to [do what we intend]." (1969:39).

That is consonant with utterances in fact having no genuine effect on our beliefs and values, at least, not a compulsory and immediate one.

For actions like stretching out for the milk, actions and conventioanl effect can be associated item by item. But this is impossible for utterances - there are infinitely many of them. There

must be a systematic relation between the utterances and the change in belief and value that it is conventionally supposed to cause. Such a connection could be provided by a compositional semantics for sentences. Such a semantics would assign to each declarative sentence (say) a proposition. But the effect of uttering that proposition would not be to make its hearer believe that proposition. Rather it would be to suggest to the hearer that he should apply plan recognition as if the utterer had tried to make him believe it.

Adopting this solution avoids Searle's point. It gives data to plan recognition which is independent of context, but does not identify that data with the proposition that is ultimately believed because the utterance was made.

1.7.1.2. Remembering what worked last time

The second possible mechanism for seeing intention in necessarily abortive actions is this: take someone who has been using language for some years. He knows, from past experience, that making an assertion has often had the effect of making his hearer accept the picture the sentence portrays. He can guess that the people he usually talks to have had the same experience. When, then, someone says to him "It is raining" he can guess that what the speaker hopes to do is to change his beliefs. "Hopes to do", not "knows he will do", because the speaker can't be sure that the regularity he has observed in the past will get him the effect that he wants this time. But this is a tremendous difference as far as plan recognition goes. Now the hearer is interpreting an action that he can believe that its performer at least thought might succeed. It is rational to attempt some actions even though they are not guaranteed to procure their end. So an action that might succeed can be explained as part of a plan, while an action that it was known must have failed can't be.

So we are entitled to suppose a speaker intended us to believe the literal meaning of his remark, because we know he knows that has sometimes been its effect in the past.

This might be a basis for explaining how some formulas get a force that they don't literally have: "Can you pass the salt?" becomes the request it is because we know it has often succeeded as such before. The first times it was encountered, a full deduction of what the speaker really intended would have been necessary. But thereafter it is treated by its hearer as having the force it had before - that of a request.

1.7.1.3. Trying to seem to try

Here is the third and last possible mechanism for seeing intention in necessarily abortive actions. When reaching for the milk, A may not be trying to reach the milk at all. He may rather be trying to seem to be trying to reach the milk. He argues to himself:

> "I cannot reach it. But let me do it in front of B. Perhaps he will argue like this:
>
> > "What is A doing? He looks as if he is reaching desperately for something. But there is nothing he can reach that he seems to want. What are the other effects of what he is doing? Aha! He has just made me think he may be trying to reach something. Could this be what he wants? If I think that, then I will think that what he wants to reach is the milk (his tea is still black). If I think that, then I will give it to him. He would want this. That must be

what he is after. I will help him"

and if B thinks that, then I will get the milk. Let me try it".

How is the understanding of an abortive reaching like the understanding of an utterance?

When A asserts something, say Fact, to B, he may argue to himself:

"I cannot make B believe Fact. But let me say it in front of B. Perhaps he will argue like this:

"What is A doing? He looks as if he is trying to make me believe Fact. But he must know he can't. What are the other effects of what he is doing? Aha! He has just made me think he may he may be trying to assert something. Could this be what he wants? If I believe Fact, then I will ((and here B will start analysing his possible responses if he come to believe Fact. If he finds a response, say Response, that he thinks A may want, he will continue ...)) do Response. A would want this. That must be what he is after. So Fact must be what he means."

and if B thinks that, then I will ((achieve the desired Response)). Let me try it".

1.7.2. Which of these is right?

Which of these is right? I don't know. I like "Knowledge of

conventional effects of actions" best. It seems to offer the best chance of systematically associating an uttered sentence with a proposition held in a form that could interact with beliefs. "Remembering what helped last time" seems too behaviouristic, and "Trying to seem to try" seems too close to founding plan recognition on plan recognition. But these are not arguments, they are guesses.

However, it would need just one of them to be true for it to be possible that, when used between sophisticated and mutually aware agents, an utterance which has no necessary consequence of changing the beliefs of its hearer may behave as if it did have such a consequence, In which case ordinary plan recognition can be applied.

1.8. How plan recognition helps communication

The problem facing the creatures trying to devise themselves a language was how to associate a content with a remark made by a speaker. Making the association either wholly dependent on, or wholly independent of, the form of the remark was useless. What they needed was a criterion by which to choose between various possible associations on different occasions.

If plan recognition is applicable to utterances, as just argued, there is such a test. It is, that since utterances are actions, they must be explicable as goal-directed. When a hearer tries a candidate association of content with an utterance, he must look ahead to the effects of accepting that content. Only associations of content that lead to an effect that is a credible goal for the speaker can be accepted.

Explaining an utterance via plan recognition can lead to changes in the hearer in a way that an utterance by itself could not. An

utterance may have a possible content C. But if a hearer attempts an explanation of the speaker wanting C to be believed, he may have to postulate the speaker's having beliefs and values that he didn't previously ascribe to the speaker. If the hearer accepts that explanation, he has to accept those postulates. Making such postulates is entirely under the control of the hearer and so isn't impossible risky in the way updates imposed by the speaker would be.

Giving a teleological explanation may involve accepting novel postulates. Accepting such postulates will involve a change of belief, value, or both. Such changes are the effects of utterances. They depend on context and are not predictable withouf plan recognition.

1.8.1. Help with choosing to accept utterances

Deciding which utterances to accept is a special case of accepting postulates about the speaker's beliefs needed to make his action explicable.

In allowing one's mind to be changed, one makes two decisions:

> that the speaker is SINCERE - that his saying something is evidence for his believing it.
>
> that the speaker is AUTHORITATIVE - that his believing something is evidence for it being true, or, equivalently, for your believing it.

Testing for deceit involves questioning sincerity. What one wants is a mechanism which, rather than making a one-off decision about the sincerity of a speaker can ask about any utterance whether it may be

deceitful. (If there were a mechanism which could detect deceit infallibly I suppose it would be redundant, since no deceit would ever be worth trying). I think one can do this by asking about speakers' purposes.

The argument that would lead one to believe a statement goes like this:

He said "X" so he intends I believe X

He would benefit (perhaps altruistically, through my benefiting) if I were to act on X and X were true

If I explain his utterance like this, I must assume he believes X is true

So he believes X is true

What leads one to guess it may be false is:

He said "X" so he intends I believe X

He would benefit if I were to act on X and X were false

If I explain his utterance like this, I must assume he believes X is false

So he believes X is false

For example:

A man whose window has just been broken by a stone rushes

out and finds two small boys, Bert and Fred. One of them will get his ear clipped. Bert cries "It was Fred!"

Bert benefits if the man believes him and and what he says is false - if it wasn't Fred - because then his own ear is not clipped. So the man has reason to suspect deceit.

The man still has to decide in this and in any other case whether he is being deceived. The merit of looking at motive like this is that he can at least distinguish cases were he need not worry about being deceived.

If you object that all this leaves no room for irony, I suggest that that "He said "X" so he intends me to believe X" is not always going to be true.

1.8.2. Help with pairing utterances and states of affairs

A a generate-and-test paradigm was proposed as a way of associating utterance and content. Though nothing that follows actually uses such a method, I include a sketch of how demanding a teleological explanation can be the test phase.

(The generate phase can rely on utterances suggesting their own content, if, for instance, the content is mediated by convention amounting to semantics. A previous section argued that that too demanded teleological explanation of utterances.)

A sentence doesn't divide possible states of affairs sharply into those of which it is true and those of which it isn't; there is a gradation. But if one knows the purpose for which the sentence is being used, the gradation is turned into a division. For instance:

A: Can the piano be moved?

B: No.

A: Not even if we get some help?

B: Oh, yes.

What is A's purpose in asking his question? He wants (B thinks) to know if the piano can be moved, presumably by A and B alone. Then it can't: he says so. But A's reply shows him A's purpose is in fact different. What now matters is whether it could be moved by several men. It could: so B says so. This sort of reason explains why A answers differently in

A: I'm worried about earthquake damage. Can the piano be moved?

B: Yes.

and

A: I've left the toddlers in the music room. Can the piano be moved?

B: No.

The relevant content of an utterance is of course identified as the only one whose uptake benefits the speaker.

1.9. Functions of different classes of sentences

Traditionally there are in English three functions that a sentence can have: statement, question, command. (There may be others, such as optative). These are supposed to let the speaker indicate what he wants done by the hearer vis-a-vis the picture of the world that the

sentence contains. The hearer should believe the world is like the picture if it is presented as a statement; he should say whether it is an accurate picture if the picture was presented as a question; and so on.

Do these functions reflect some fundamental limit in the world on the things that one can do with a picture? If they do, then whatever they have in common ought to suggest what it was that made utterances into effective actions. For instance (to take an absurd case) it might be that we were neurologically incapable of not formulating an answer to a question or an obedient action in response to a command. Then the effect of utterances should be explained by our physiology. Alternatively, the functions might be reducible to some more primitive operations. If so, these primitive operations might suggest the sort of thing that they could operate on; and that would suggest what sort of thing one would have to look at if one wanted to predict the effect of some particular utterance.

There seem to be two obvious ways of analyzing function: One can think of commands as basic. Then

- a command to do X is a command to do X
- a question whether X is a command to say whether X
- a statement that X is a command to believe that X

Or one can think of statements as basic. Then, if one supposes that one can get other friendly people to do what you want by telling them what it is that you want;

- a command to do X is a statement that Sp wants Hr to do X
- a question whether X is a statement that Sp wants Hr to say whether X

- a statement that X is a statement that X

How can one choose between these? One example that may show that the second is preferable is this:

L and S have just got up. They set the washing machine going the night before, but it has been being temperamental recently. Neither has yet seen it this morning.

L: Has the washing machine worked?

Now why did L ask? She knows that neither of them knows the answer. If asking questions is best seen as giving commands, then L's question has to be seen as either an impossible command, because S is known not to know the answer, or else a command, not just to say whether the washing machine had worked, but to go and find out and then bring back a report. The trouble with the second analysis is that L wasn't, as a matter of fact, in the sort of social position that would make such a command a reasonable one. It involves too much trouble to be a reasonable request.

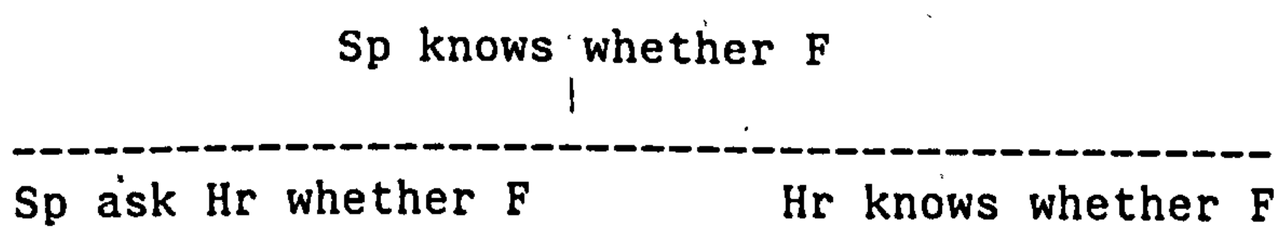
This need not be a fatal objection. One could say that the question was a command, but one which L expected, even intended, to fail. But if S asks the question "What was L trying to do when she gave that command?" he is going to have to suppose that she wanted to know the answer. Even if he doesn't do his part to make L's plan work, he is left with this inference, and may fulfill L's want later. If S can suppose that L expected that, he has an explanation of what L said.

There may be an analysis that both explains the command-like feeling that questions generate, while not committing one to taking commands as primitive. It runs like this. The speakers of the language know

that the interrogative marking of sentences is used in the expectation that people hearing it will reply to the question. Where this expectation comes from doesn't matter. Nor need it be a very certain expectation. It may be a piece of knowledge rather like what we know about 30 mph speed limit signs. We all "know" that they make people drive slower, even if the knowledge that we use when crossing the road as pedestrians contradicts this.

1.9.1. How do questions work?

If this is so, one could understand most questions as part of a plan like this:



But this is too compressed. It doesn't show that Hr has to reply in order for this scheme to work. Nor can it explain this case: A psychiatrist wants to find out whether an old person is oriented for time and place. So the psychiatrist asks "Is this March?". The old person's answer is not going to have any effect on the psychiatrist's beliefs about whether it is March or not. But the schema above suggests that it should. The point is that a question may provoke a reply, but a reply need not provoke belief. I must make this contingency explicit.

Before you can believe in an answer to a question, you have to be sure of two things:

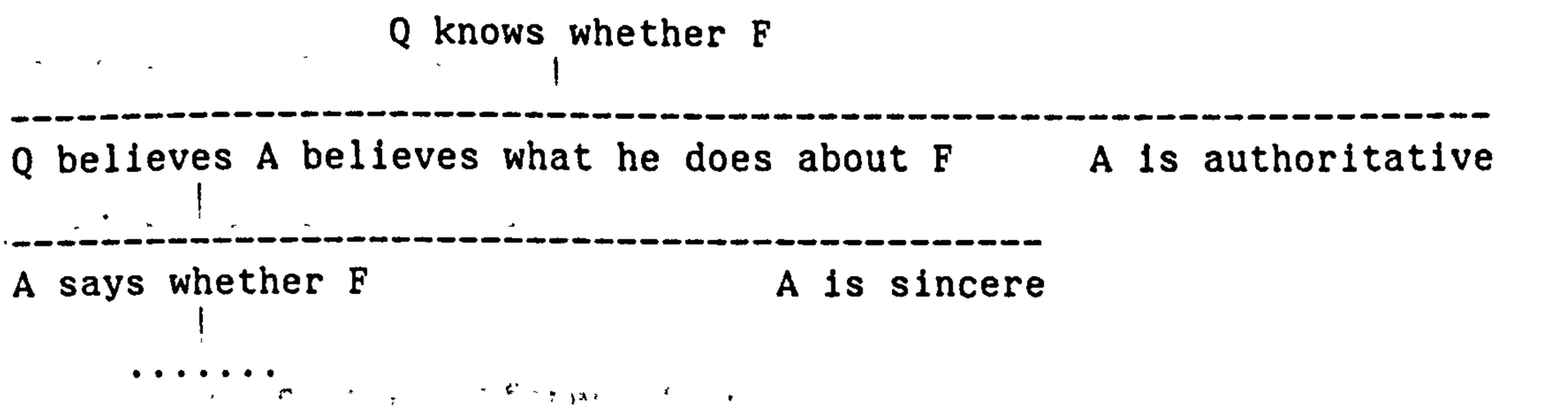
- The answerer must be SINCERE. If he says something, this must be because he believes it.

- The answerer must be AUTHORITATIVE. If he believes something, it

must be good evidence that what he believes is true.

These are obviously separate conditions. A man whose answer is authoritative may be insincere: for instance, a thief who denies that he stole. A man whose answer is sincere need not be authoriative: as when you on boarding a train ask me, already on it, if the train goes to Gloucester, and I say it is, with the result that we both inadvertently go to Exeter.

To accommodate these extra conditions, perhaps the final stages of the plan that a question is part of look like this: (Q is the questioner. A is the answerer.)



On this account, saying that someone is sincere amounts to accepting the rule

A says X -> A believes X

and saying that A is authoritative amounts to accepting the rule

A believes X -> X

In fact one would only accept someone's word on certain matters. A better rule would be

A believes X & X is a fact that A is likely to be right about -> X

but the principle is the same.

The next question is, how is A brought to say anything? One might suppose this rule would be sensible:

A says whether F
|

Q asks A whether F A knows whether F

I have used expressions like "know whether", "say whether" as if they represented attitudes to, and actions on, proposition which were of the same sort as, even if different from, knowing and saying. This is I think wrong. To say that someone knows whether X seems to be same as saying that either he knows that X is true or that he knows that X is false. In symbols

A knows whether F = A knows F v A knows -F

which is of course different from

A knows whether F = A knows (F v -F)

But then what about the states in the plans above that refer to "knowing whether"? How can one expand them? The obvious thing to do, which is to substitute into the plans using the equivalence above produces nonsense. Here is the plan with the substitutions:

Q knows F v Q knows -F
|

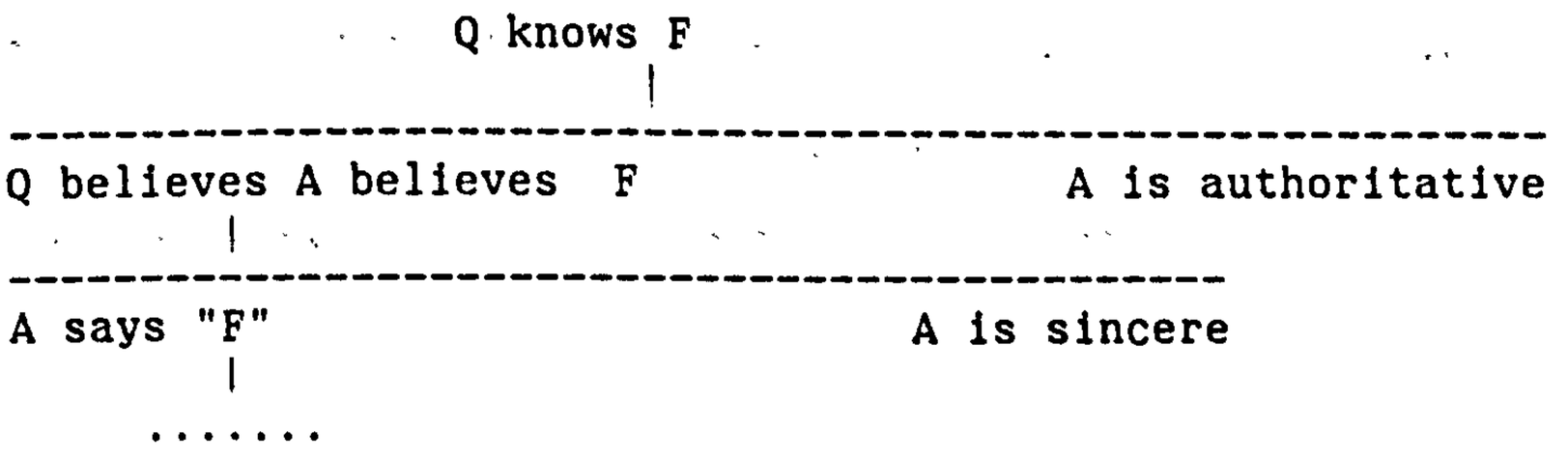
Q believes (A believes F v A believes -F) A is authoritative
|

A says F v A says -F A is sincere
|
.....

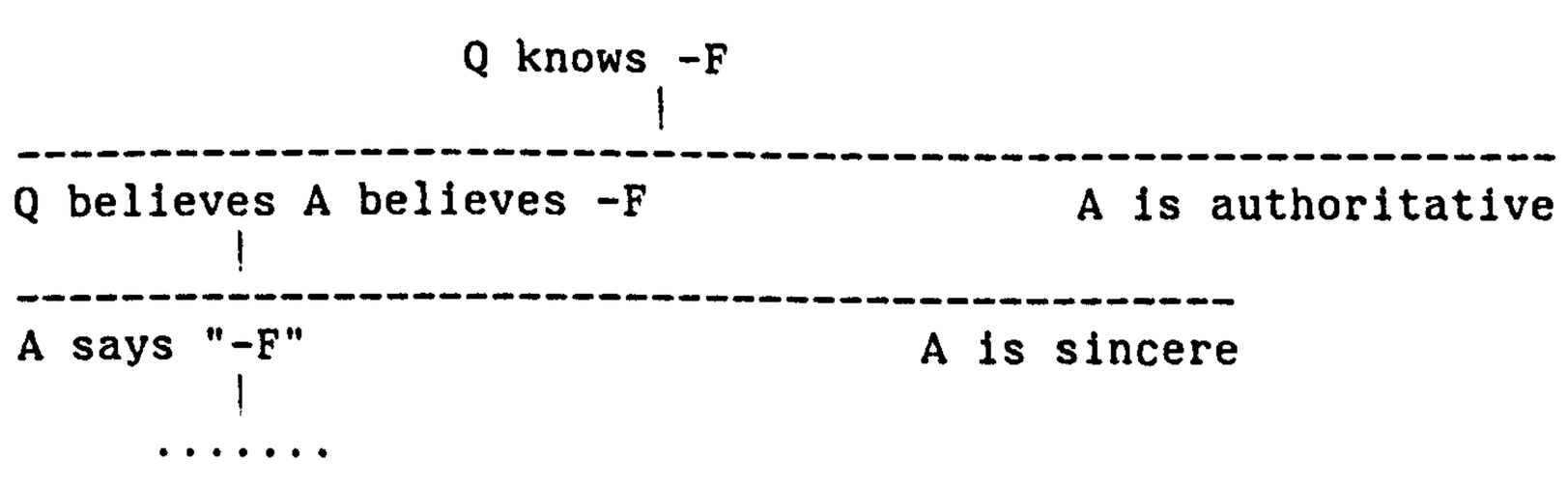
The top sections of this appears to say that if Q believes A is authoritative and if he also believes that A knows the answer (that

is, either believes F or -F, and is not just in doubt), then Q too can be certain about F. This must be wrong. I may well know that you know whether your front door is red, and accept that you are authoritative about it, but still not know whether it is red or not. Something is wrong.

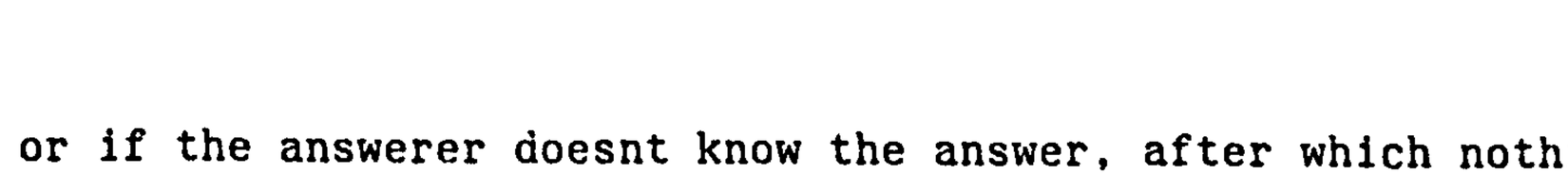
What one is really doing when asking a question is performing an essentially risky action that may have one of several outcomes. The question provkes the answerer to reply. He may say "yes" or "no" or else admit he doesn't know, or refuse to answer. You can't be certain what will happen. Different outcomes will follow your question, depending on how the world was when you asked it. The outcome may be this, if the answerer says "yes":



or else this, if the answerer says "no"



or if the answerer doesnt know the answer, after which nothing can be deduced



The problem with the proof tree I complained about above is that it

is a comprehensible but mistaken attempt to conflate the trees below it.

From this I conclude that there is no need to have any ingredient to explain the effects of different sentence functions which is not an appeal to planned, though perhaps risky, action.

1.10. Conclusion

It seems that there are in discourse several phenomena that cannot be properly described without looking at the plans of the people talking. Indeed any account in which speakers are not able to coerce their hearers' beliefs seems to involve planning. But to say anything more detailed about what constitutes such discourse events, more must be said about how plans are made and changed. After looking at related work, I shall attempt to fill that need.

Chapter 2. Related work

This chapter looks at related work. The first part is about the philosophical background in the work of Austin, Grice and Searle, and the work of Gordon and Lakoff, on the boundary of philosophy and linguistics. Then I consider two ways of integrating utterances into discourse:

- Allen's, which understands utterances by guessing the goals they serve and supposing that the goals persist throughout the discourse

- The "Yale school" view which sees actions (and so utterances) as fillers of culturally determined scripts.

2.1. Austin

The observations about speech acts that started the continuing current interest in them were made by JL Austin, apparently around 1939. They were the substance of his 1955 William James lectures at Harvard, which were later published as (Austin 1962).

Austin observed that utterances could be divided into two (non-exclusive) types: those that could be true or false - "constative" and those we would usually say were neither, but which instead are such that, when one makes them, one is doing something. These he called "performative":

Austin's way of understanding the force of a performative relies very heavily on there being some "conventional procedure, having a certain conventional effect, that procedure to include the uttering of certain words by certain persons in certain circumstances" (1962:14).

An utterance is a performative if it is an instance of such an utterance. Clearly it is possible to make an utterance of the correct form even outside the correct circumstances. When that happens the utterance is guilty of "infelicity". The required circumstances can include states of intention and belief on the part of the speaker, and it is by detecting infelicities there that he explains how a promise can seem to be made by a man who uses the words "I promise ..." although he has no intention to perform.

Austin points out that there are utterances that are both constative and performative. One such class is those like "I blame ..." which is both a conventional act ascribing the social quality blame, and a report of an internal state. Another class is those that he calls "expositives", such as "argue", "conclude", "testify", where it is impossible to do the conventional performative act without also doing the related constative. Ultimately he seems to treat constatives as performatives (describable as eg "stating", "maintaining") that as well as performing that act, can also be true or false.

Separarate from the performative/constative distinction, Austin makes another which has been even more influential: it is that between the locutionary, illocutionary and perlocutionary acts (or locution, illocution and perlocution) involved in an utterance. The locution is "the utterance of certain noises [...] with a certain meaning. [...] ie with a certain sense and a certain reference" (1962:94). The illocution is "the performance of an act IN saying something, as opposed to the performance of an act OF saying something" (1962:99), which he also glosses as the force that certain words have. The perlocution is "[the producing of] certain consequential effects upon the feelings, thoughts or actions of [...] other persons; and it may be done with the design, intention or purpose of producing them." (1962:101). Both the locution and the illocution involve convention:

the perlocution does not.

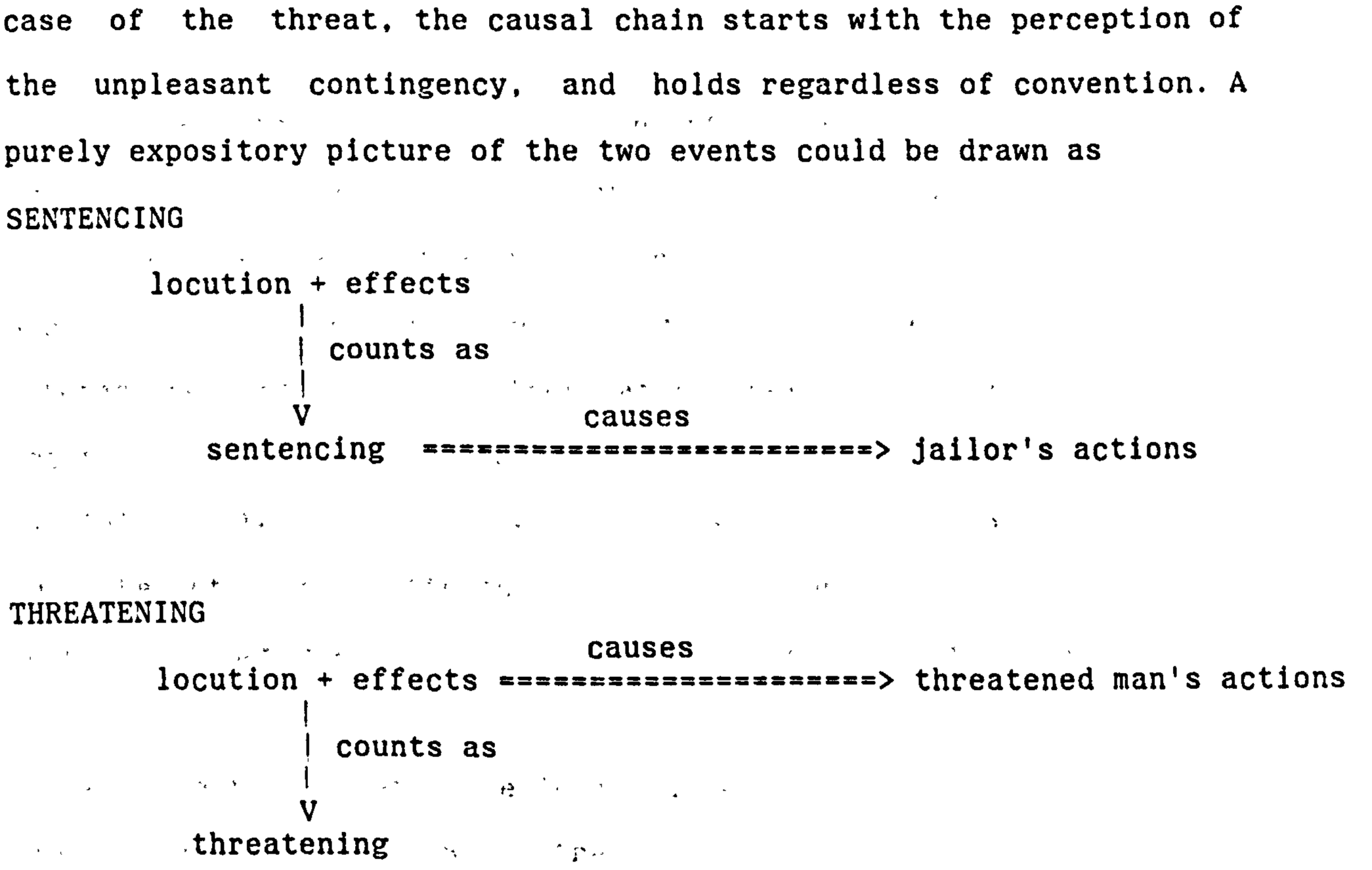
This distinction involves him in having to separate those things that are true as result of an action (an utterance) because of what the action is, and those that are true as a result of what the action causes. The first are ascribed to the illocution, the second, to the perlocution. He takes three chapters to do this (IX-XI) and they are much fiddlier and less compelling than what has gone before. This is (I claim) the result of his system giving illocution an explanatory load greater than it can bear. It is the essential step that determines which speech act is being performed. All utterances have the force they do because of their illocution. Illocution is constituted by convention. So utterances have force because of convention.

Instead, I suggest, we should split what Austin attempts to explain by illocution in two. There are utterances, such as baptism, sentencing a prisoner and so forth for which his model seems wholly correct. A judge's sentence has its effect because it is recognized by jailers for what it is; and to recognize it, an understanding of certain conventions is essential.

On the other hand, the supposed illocution in "threaten" (1962:121) is different. What constitutes a threatening is (roughly) making someone aware that if and only if they do some action, then the threatener will do some other action which is harmful to the person threatened. The perlocution of a threat will then follow as the response of any rational agent to discovering an unpleasant contingency of a contemplated action.

These two cases are alike in this: the judge and the threatener have both done some things and caused others. It is a result of convention

that the judge's action and its effects are an instance of the definiens of the word "sentencing", and that those of of the threatener, of the word "threatening". So far, I follow Austin. However, the jailer's response only occurs because he knows how in his society he should respond to sentencings. His grasp of convention is part of a causal chain leading to his future action. But in the case of the threat, the causal chain starts with the perception of the unpleasant contingency, and holds regardless of convention. A purely expository picture of the two events could be drawn as



If this is accepted, then a revision of Austin would be possible: the illocution of an utterance is to be whatever that utterance is called in the lexicon of that society. To describe people's responses to an utterance, one will need laws describing how awareness of its being called that is causal in the behaviour of its hearer. But with such an (admittedly large) modification, it seems to me that Austin stands.

I follow Austin in seeing utterances as acts having effects, and therefore falling under certain descriptions. The problem with his account is, not that he does not state what the effects of each

utterance type are, but that he does not even describe in detail what sort of effect they might be; and that he does not explain how context affects the effect of an utterance. I attempt very partial answers to these; effects are to be effects on mutually-known proof trees, and (part of) context is to be the set of such trees.

2.2. Grice

In his paper "Meaning" (1957), Grice introduced a distinction between two sorts of meaning, and showed, by ordinary-language arguments, that they behaved differently. One he called "natural meaning". It occurs in such usage "Those clouds meant rain". The other, more interesting sort, he called "non-natural meaning" or "meaning-NN". It occurs in such usage as "Those three rings on the bell meant the bus is full". The sort of thing that means-NN is an action by a person, and the interesting effects of such an action are effects on another person. The paper analyses actions that mean-NN and their effects.

Grice rejects, first, the idea that such an action has its effects via a causal chain in which the hearer's beliefs are altered willy-nilly. Then he goes on to make the first of two vital steps; he wonders if the hearer's recognition of the speaker's intention is central. He rejects the simplest notion, that meaning-NN consists in having an intention to induce in the hearer both a certain belief and belief in that intention. Instead, he proposes a more complex notion, which he summarizes in (Grice 1968) as "[Speaker] intends to produce in [Hearer] effect E by means of [Hearer's] recognition of that intention". The effect is to be more general than just a change in belief. The second vital step was the introduction of the reflexiveness in an intention that that same intention be recognized.

The paper concludes by disavowing the idea that when talking we

explicitly consider or recognize other peoples intentions, and roughing in an alternative in which we rely on our grasp of normal usage of words and expressions to abolish the need for so much folk psychology. So, where Austin sees the force of utterances mediated by the conventional action they embody, Grice sees it mediated by reflexive recognition of intention. Grice does however rely on conventional usage to explain how we recognize from the speaker's action what his intention is.

The trouble is that one still needs to know how grasp of convention affects recognition of intention, and the roughing-in is not satisfactory. Grice attempts an elaboration in (Grice 1968). There he lays out for himself a programme of defining and distinguishing the sort of thing that is the meaning of a speaker's utterance (construed widely) on a particular occasion ("speaker meaning"), and the meaning that the linguistic machinery used to make that utterance conventionally bears ("utterance meaning").

His account of speaker meaning is essentially that of (1957), except that he takes the intended effect of an indicative-type utterance to be, not that the hearer believes something, but that the hearer believes that the speaker believes something; and the intended effect of an imperative-type utterance to be, not that the hearer does something, but that the hearer intends to do something.

That seems eminently right. But his account, on his own admission, of explaining utterance meaning is defective: "So I shall for the present abandon the attempt to provide a definition, and content myself with a few informal remarks.". I think it fair to state his final position as being: An utterance type U has conventional (utterance) meaning E for Speaker if Speaker is able to follow a group's (or his own) practice of saying "U" when he wants to have

effect E on his hearer.

He concludes the paper with a complex attempt to relate the compositionalities of speaker meaning and sentence meaning. I mention this because Searle's interest in speech acts seems to start in the same area.

Grice also elaborates the notion of non-conventional meaning-NN. Early in (Grice 1968) he declares he is engaged in trying to describe the total significance of a remark, part of which has been "said", but part of which has been "implicated" (= "in some sense, implied"). What is implicated can be implicated either conventionally or non-conventionally. He develops the latter further in (Grice 1975), originally delivered as lectures in 1967.

His initial observation is that in the exchange where A asks B how their friend C, now working in a bank, is doing, and B replies "Quite well; [...] he hasn't been to prison yet", B is strongly suggesting, "implicating", that C is dishonest. How can this happen, given that such an implication is no part of any conventional meaning in the exchange? He calls it a "conversational implicature", in contrast to a "conventional implicature", that might be part of the conventional meaning. Grice points out that this implicature is not an implication, since the statement could be true and the implicature false.

His solution has this structure: suppose that there are certain constraints on discourse. Suppose, when these are violated by some remark P, a hearer will (and will be expected to) postulate the truth of another proposition Q, such that if Q is deemed to have been said (as part of P?), the constraints will be satisfied. Then P is said to have "conversationally implicated" Q.

Since one has to know the context to decide if some candidate Q is reasonable, Q cannot be a conventional part of P. Further, since there is a choice of Qs, the falsity of any one will not imply the falsity of P, so the implicature is not an implication.

Grice's choice for the role of the constraint in this process he calls the cooperative principle: it is, roughly, "make your conversational contribution such as is required [...] by the [...] talk exchange in which you are engaged". This is expanded as a set of maxims to be obeyed - maxims of quantity, quality, relation and manner. He gives examples of how different ways of failing to fulfill these all signal the need to start looking for a conversational implicature.

I shall argue later points similar in structure to Grice's, though different in content. I shall argue that insisting on the mutual recognition by the speaker and hearer of each others' intentions is right - but I claim that the intention that has to be so recognized need not be the intention to communicate; instead it is whatever intention it serves their purposes to talk about. I shall also argue that postulating an unspoken proposition whenever the discourse seems to need one to fit constraints is also right - but that the constraint is really that each piece of the discourse should interact with already mutually-known proof trees in a certain goal-advancing way.

2.3. Searle

In Searle's book "Speech Acts" (Searle 1969) his main interest is analysing reference and predication by analysing their relation to the speech acts in which they occur. For him, "a study of the

meanings of sentences is not in principle different from a study of speech acts" (1969:18), and complete speech acts are illocutionary acts (1969:23). I shall however consider only his discussion of illocution and speech acts.

An utterance usually involves both a propositional and an illocutionary part. The propositional part is, unsurprisingly, tightly bound up with the knowledge and observance of conventional rules by the speaker and hearer: those governing the language used, for instance. But illocution too depends on the existence of rules, and Searle is more detailed than Grice in describing this dependence. Perlocutions exist too, but are not constitutive of a speech act, though they may be mentioned in the definition of a speech act verb.

Searle's counterpart of Grice's meaning-NN is the act of saying something and meaning it. He defines the action like this:

S means "what he says" by the utterance U to H of sentence T iff

(i) S intends U to produce in H the effect E, where E is the belief that the state of affairs conventionally expressed by T is true.

(ii) S intends U to produce E by H recognizing that (i).

(iii) S intends that (i) will be recognized because H knows rules governing T.

But this action is an illocutionary act. So because of the novel last clause, Searle makes illocution depend on convention as well as on intention.

The same framework can be used to define other illocutions. But the details are quite tricky. The point is that one can give conditions under which an utterance does constitute the illocutionary act of promising, (and Searle does so, (1969:57ff)), but these rules are not the sort of rules referred to in (iii) above, though they will be closely related. The first sort of rule is about when an utterance is a promise; the second, about how an utterance may announce itself to be a promise.

An utterance announces itself to have a particular illocution by satisfying "felicity" conditions of four types: the preparatory, sincerity, propositional content, and essential conditions. These may be conditions on either the utterance or the content of the utterance. Searle gives many examples of these for different illocutions. The preparatory condition for for instance a directive (a request) is given in (Searle 1975) as "Hearer is able to perform [the requested act]".

In (Searle 1975) he tackles the problem of utterances where the illocution clearly present is different from the illocution that ought, according to convention, to be there. For instance, the evergreen example "Can you pass the salt?", which should be a question but is really a request. Such cases, where one speech act is performed by performing a different one, he calls indirect speech acts.

Responding to such utterances is to be a two-stage process. First one must spot that an indirect speech act has occurred: this is to be done by seeing an apparent violation of something like Grice's principle of cooperation. Second, one has to find which indirect speech act it was.

To do this, he re-uses the four conditions mentioned above. (Though interestingly, and I suspect inadvertently, they are here taken as "conditions that are necessary for the successful and felicitous performance of that [illocutionary] act" in an utterance (1975:44), rather than rules about how an utterance may indicate its illocution; but it doesn't matter.)

The general scheme is that an indirect speech act of some kind K can be made by asking whether, or stating that, (or sometimes either), any of the four conditions on a direct speech act of kind K is true. For instance, the preparatory condition on a request is that the hearer is able to perform the requested act: so "Please pass the salt" only announces itself to be a request if my hearer can pass the salt. But if I were to state that condition (in "You could pass me the salt"), or to question it, (in "Can you pass me the salt?"), I could be making an indirect request. Again, Searle provides many examples.

My objection to Searle's approach is three-fold.

Firstly, I think the definition of illocution in terms of the four felicity conditions is flawed. The parallel between conditions for the performance of a speech act, and indirect ways of performing the same act is not smooth. The "essential condition" (which for instance for directives is that the utterance "counts as an attempt by Speaker to get Hearer to do Action") might be expected to sustain indirect speech acts just as the other conditions do. But one cannot make indirect requests by stating that or asking whether the essential condition obtains, as with the other conditions. Instead, one does it (usually) by stating that, or asking whether, there are good reasons for doing the act. I think this is a symptom of a deep problem. Searle sees that the fact that an act satisfies the preparatory,

sincerity, and propositional conditions is not enough for it to have to be the appropriate illocutionary act, and so adds the essential condition. This, unlike the others, is non-operational. Since the essential condition is not operational, questions or statements about it aren't operational either, and so they can't be used to mark indirect speech acts.

I claim that the reason that he doesn't give an account of how one can tell whether the essential condition is fulfilled is that it would be impossible; the condition is ill-defined unless one looks at how speaker's and hearer's plans change. Indeed, if it was possible to define it, it would presumably subsume the other conditions. How could an speech act count as a directive if it demanded an action that the hearer couldn't do and so violated the preparatory condition?

Secondly, I want to see speech acts defined by intended perlocution rather than illocution. The relevant perlocutionary effects will be either effects on mutually-known proof trees (and perhaps on what Searle calls "institutional facts" - facts about our social environment - which I would need to explain such speech acts as sentencing). If I can, I will not need clause (iii) of Searle's revision of meaning-NN, nor the complex but still wholly pre-formal machinery he needs to implement it.

Further, I guess (I cannot show) that were I to meet a creature that spoke my language but which came from an arbitrarily different society (a Martian, for the sake of argument), we would still be able to make indirect requests to each other. If it said "Have you got a space helmet?" I could reasonably take this to be a request for a space helmet. If this is so, then either knowledge of the conventions necessary to sustain illocution is a part of knowledge of a language;

or speech acts should be explained without reference to such convention. This is not an argument for my position or against Searle's. It is an argument for why my position should be examined to see if it is feasible.

Searle explicitly rejects this sort of account founded on intended perlocution when criticizing Grice. He argues

> "When I say 'Hello' and mean it, I do not intend to produce or elicit any state or action in my hearer other than the knowledge that he is being greeted. But that knowledge is simply his understanding what I said, it is not an additional response or effect." (1969:46)

and again

> "[...] I may say something and mean it without in fact intending to produce that effect. Thus I may make a statement without caring whether my audience believes it or not, but simply because I feel it my duty to make it." (1969:46)

But the first seems flat contradiction; change of knowledge is an effect. And the second misunderstands what the intended perlocution is: there, the perlocution is the bringing about of the (institutional) fact that on this occasion the speaker stood up for his principles.

My third objection is that Searle's approach underplays something that he does mention. In sketching the process of finding which indirect speech act has been performed, Searle imagines the hearer testing the candidate speech act to see whether it is one the speaker

might have wanted to make. In the salt-passing example, the hearer is imagined as thinking that because he and the speaker are at dinner, getting the salt is something the speaker might want. This testing is however appqrently merely a heuristic. It chooses the best from from a candidate set defined by rules about questioning and stating the illocution conditions. I want to argue that in fact this teleological test is central, and that 'being useful to the speaker' is the essential property that defines which speech act is performed in an utterance. My reason for this is that 'being useful to the speaker' is the only credible convention-free criterion available.

2.4. Gordon and Lakoff

Gordon and Lakoff (1975) attempt two things: to make more formal Grice's intuitions about how non-literal communication arises from general principles of conversation; and to show that syntax is sensitive to such principles. I consider only the first part.

Their central machinery is deductive. The "conversational implications" of an utterance are to be the ordinary implications that follow from the conjunction of its literal meaning, a description of its . context, and certain extra axioms, the "conversational postulates".

The set of conversational postulates does not seem a priori to be closed. But interestingly they attempt to derive some of them systematically. They observe that there are conditions on the sincerity of requests that can be formalized to give sentences such as

i SINCERE(a,REQUEST(a,b,Q)) -> WANT(a,Q)

ii SINCERE(a,REQUEST(a,b,Q)) -> ASSUME(a,CAN(b,Q))

i is called "speaker-based" and ii, "hearer-based", because of the distribution of person-variables in the consequent. These principles are not themselves conversational postulates, but axioms about rational behaviour.

They add a principle (called (6)) that "one can convey a request by (a) asserting a speaker-based sincerity condition or (b) questioning a hearer-based sincerity condition" and so derive (with an explicit fudge about ASSUME)

i' SAY(a,b,WANT(a,Q)) -> REQUEST(a,b,Q))

ii' ASK(a,b,CAN(b,Q)) -> REQUEST(a,b,Q))

i' and ii' are conversational postulates. When someone says "Can you take out the garbage?", the antecedent of ii' becomes true, and the presence of the indirect request is revealed by simple deduction.

They also have a notion of "challenges" of unreasonable speech acts. A challenge to a request occurs in such remarks as

"Why do you want me to do that?"

"I can't do that - I hurt my arm"

They offer a similar general principle, (called (14)), which can be re-expressed as the function from sentences to sentences

SINCERE(a,F) -> C

=>

REASONABLE(a,F) -> (Er)REASON(r,C)

for F of the right form. From this and the same sincerity conditions used before they can obtain

i'! REASONABLE(a,REQUEST(a,b,Q)) ->

(Er)REASON(r,WANT(a,Q))

ii'' REASONABLE(a,REQUEST(a,b,Q)) ->

(Er)REASON(r,ASSUME(a(CAN(b,Q))))

It is because of these that the challenges above are supposed to have their force. Both a question about S and a denial of S tend to suggest -S. The question presumably suggests -S because asking it presupposes that S is not provable, though this is not made clear. One challenge asks about the reason for a want, the other denies any reason for assuming the requestee's ability. So by modus tollens each implies the negation of the reasonableness of the request.

Another example, which is not connected to sincerity conditions on actions, is an attempt to explain why the unprovoked remark "Your wife is faithful" is insulting. Tidied up, their account assumes a conversational postulate

SAY(a,b,Q) -> INTEND(a,KNOW(b,Q))

and a general rule

INTEND(a,X) -> ASSUME(a,-X)

so that one can derive

SAY(a,b,Q) -> ASSUME(a,-KNOW(b,Q))

Hence the remark "Your wife is faithful" entails that the speaker believes his hearer does not know this - commonly an insult.

There seem to be two problems with this account. The first is that it seems to have made the wrong things explanatory. The postulates show

how to derive REQUEST(a,b,Q) from SAY(a,b,WANT(Q)); and clearly sometimes this will be very useful. One can only characterize the action "rudely ignoring a request" if one has some prior definition of "request". But ignoring such institutional facts, what is there that is needed to predict how my actions will change which is included in saying that someone has made a request of me, but which is omitted from saying that they have expressed their wants to me? Two possible answers are, that I cooperate with the requester, and that the requester was sincere. Those points seem right. But Gordon and Lakoff's account covers neither; it does not mention cooperation at all, at least in the account of why "Can you take out the garbage?" is a request; and it does not suggest any way one might test whether an ostensible request is sincere. (I would like to have claimed that for them the proof of the occurrence of a request depended on prior proof of the sincerity of any utterance, but I can't. The relation between i and i' is not implication, but rather a sort of meta-rule-like projection.)

The other problem is, where do the conversational postulates come from? Where do principles like (6) and (14) come from? They have a lot of axioms. I shall try and argue that the regularities they see are real, but are in fact consequences of actions that alter what is mutually known about the participants' plans. For instance, that "asserting a speaker-based sincerity condition" (that the speaker wants Q or believes -Q) is a request is not an axiom, but a consequence of the fact that if it becomes known that Sp wants Q (when it can be assumed that he believes that -Q) or if it becomes known that he believes that -Q (when it can be assumed that he wants Q) then from general principles of rational action it will follow that he will want actions by his hearer that tend to achieve Q. However, other facts too will follow (for instance that he will want actions by himself that tend to achieve Q) and this will explain

indirect meanings (such as requests that his hearer abstain from actions that would prevent his own) that Gordon and Lakoff do not contemplate at all.

Gordon and Lakoff's approach has been continued by Gretchen Brown (1980), but with a rather different purpose. She accepts that an approach such as Allen, Perrault and Cohen's may be the general explanation of speech acts, but argues that what Gordon and Lakoff propose is an "appropriate technology" for recognizing indirect speech acts in discourse, and so presumably for systems in which natural language communication is to be used rather than studied. Once she has claimed to be giving a purely descriptive account, it is not weakened by the fact that she has to list literally dozens of postulates. There is still a difficulty in explaining how one can recognize the antecedent of one of the postulates in an unprocessed input sentence. Eg her rule NECESSARY-ASSERT says that one can perform a speech act by asserting that it is necessary to perform that speech act. But how exactly is "I must ask you to join the Marines" an "assertion that the intended speech act is necessary" and so a way of making a request? And a purely deductive NECESSARY-ASSERT postulate should say that if one asserts that it is necessary to perform a speech act, then one DOES perform that speech act (though that problem can I think be patched). Within the limited scope Brown considers, the Gordon/Lakoff/Brown approach seems extremely practical.

2.5. Allen's work

2.5.1. Background to Allen

Work in which planning and natural language meshed has been done for some while (eg Grosz 1977), but the first that mixed this with the

philosophical tradition mentioned above came from Toronto, where Cohen, in his thesis (1978), and Perrault and Cohen (1979), wrote about the production of speech acts by a planning system. This relied heavily on the notion of recursive belief and Searle's analysis of speech acts. It also used action schemata with preconditions on the wants of the action's agents.

Shortly afterwards, Allen wrote his thesis (1979), about the recognition of speech acts rather than their production, which relied on the notion of recognition of intention. Allen's ideas were also used by a group at BBN (Brachman 1979).

Later work in the tradition has been that by Perrault and Cohen (1981) on recursive belief and what speakers do when they refer; and the ARGOT project under Allen at Rochester (Allen 1982, Allen et al 1983), still in its early stages. The most striking change there to my eye is that the acts performed in discourse are not all at one level. Some are as always concerned with the state of the domain, but others, such as attempts to change topic, act on the discourse itself.

Below I concentrate on Allen's ideas. They occur in many papers, but the one that I find the clearest and the most readable is one by Allen (1983). He himself describes it as his final attempt to clean up his thesis work.

2.5.2. Allen and the recognition of intention

Allen considers a system to which the user can say something - often but not always a question. His system imitates an information clerk at a station, and the user, a man meeting or catching a train. From what the user says, the system deduces what plan the user has. Then

it looks to see what obstacles there are in this plan, and attempts to remove them. These obstacles are lacks of knowledge which can be removed by giving the user information.

(In the work done at BBN, the system actually went further and achieved goals for the user. The system could display different parts of a ATN on a VDU in different ways, and would change what was displayed to accommodate the user's indirectly expressed goals.)

If the system deduces the obstacle from the utterance, then the user's utterance has had the indirect effect of revealing the obstacle. So it is an indirect form of whatever direct speech act would have revealed the obstacle. So the system can recognize indirect speech acts.

Recognition of the user's plan is done like this: The system starts with the observed utterance and some expectations about the user's goals. The utterance is taken to be an action directed to some yet unknown goal, which will probaby turn out to be among the expected goals. The problem is to fill in the gap between the observed action and the expected goal. This gap can be closed from either end. One can work down from the expected goal - if the user wants to catch a train, then the system can infer that he probably wants to be at the platform it leaves from. This is an instance of a general "plan construction" rule, that "If a person wants to do some act, he probably wants the preconditions of that".

Some of the rules refer, not to domain-specific actions or states that will occur in the plan, but to states of the planner's knowledge. He may have the goal of knowing the reference of terms such as "the departure time", or whether a train has arrived yet.

But the system can also work the other way. There are inference rules that allow the system to infer from goals that it already thinks the user has to other goals that generated them. These are "plan recognition" rules. They are the inverses of the rules that one might use to construct a plan. For instance, the inverse of the plan construction rule would be "If someone wants the precondition of an action, then perhaps he wants to do the action". Allen has a list of about a dozen such rules.

The limiting case of a goal whose ulterior purpose has to be recognized is an observed utterance. This, and any partial plan deduced from such by the plan inference rules, but which has not got far enough to join up with an expected goal, Allen calls an "alternative".

Several rules may be applicable to one alternative, so from one alternative it may be possible to guess many others. Allen's system has to know what it should infer, as well as what it may. He does this by using two sorts of numerical weight. One is attached to the alternatives; the other, to tasks. There are certain actions that can be performed on an alternative: for instance, applying one of the plan construction or plan recognition rules, or deciding which can apply, or accepting an alternative as so much better than its competitors that it must be the user's real goal. A task is the opportunity of applying one of these actions to an alternative. The merit of each task is a function of the type of the task and the merit of the alternative that it is to be applied to. The best is done. Ultimately the task of accepting one fragment as definitive is the best, and when it is done the process stops. The merit of an alternative depends on how well it matches an expected goal (eg do they refer to the same objects?), on whether it relies on things known to be possible, and on whether inference rules have been

fruitfully applied to it in the past.

What the system eventually produces is something like this:

```
BOARD(A,train1,TORONTO)
        |                                          enable
AT(A,loc1,time1)
        |                                          know
KNOWREF(A,time1)
        |                                          effect
INFORMREF(S,A,time1)
        |                                          want-effect
WANT( S, INFORMREF(S,A,time1) )
        |                                          effect
REQUEST( A, S, INFORMREF( S, A, time1) )
```

I have described all this without referring to recursive belief or belief contexts. Let me take one of Allen's rules as he gives it (though I have renamed the agents):

Know Positive:

```
Hr b Sp w (Sp KNOWIF P)
-----------------------
Hr b Sp w P
```

The initial modality of the sentence means of course "hearer believes speaker wants ...". All the sentences that start with the same modality are said to be in the same belief or want context.

If you want to see what was actually deduced during the building of the structure above, prefix each line in it with "Hr b Sp w ... ", since it is in that context that simple plan recognition is done.

The sentences on the main line are goals of the user. The words on the right are the names of the inference rules employed, (or at any rate a clue to why the link is allowed).

But though Allen calls it a plan, it isn't. States resulting from actions (AT) are mixed up with domain actions (BOARD). Knowledge goals (KNOWREF) are included as if they were ordinary preconds of an action. And the whole thing is linear. No goal can have more than one sub-goal. What it is of course is the trace of a series of inferences about a plan, rather than a plan. As a result, all the nodes in it are things made desirable by the plan.

So much for guessing the speaker's intention. But Allen wants his system to go on and make a cooperative response. The next stage is to detect obstacles to the speaker performing his plan. For instance, catching a train requires one to know where and when it leaves. If the plan that the system ascribes to the speaker involves this, and if further it knows that he doesn't have this information, then it can co-operate by giving it.

I think this is one place where not really recognizing plans may hurt. Suppose you ring the station and ask the information clerk when the Toronto train leaves. Unless he makes a guess at your full plan, eg that you are going to come to the station by taxi, he won't ever think to tell you about such obstacles as all the taxi drivers being on strike.

That is roughly the machinery needed for making co-operative responses. But as Allen observes it can run without having to consider whether the speaker was actually attempting to get the co-operative response. Allen identifies such an attempt with the performance of an indirect speech act. How can such attempts be spotted?

He uses two extra notions:

- mutual belief. A believes that A and B mutually believe that P (written as mb(A,B,P)) if AbP & AbBbP & AbBbAbP & ...., where "b" stands for "believes"

- a distinction between illocutionary, "real", REQUEST and INFORM acts, and SURFACE-REQUEST and SURFACE-INFORM acts, which are just the utterance of sentences grammatically-marked in the right way. The definitions for the requests are

REQUEST(speaker, hearer, action)

body: MB(hearer, speaker, speaker WANT hearer DO action)

effect: hearer WANT hearer DO action

SURFACE-REQUEST(speaker, hearer, action)

effect: MB(hearer, speaker, speaker WANT hearer DO action)

The intuition behind this is that if what constitutes the body of the act arises then ipso facto the act has been done. Any act, such as a SURFACE-REQUEST, is a "real" request just in case it makes the body of REQUEST occur.

One problem here is that Allen suggests that the preconds/body/effect structure of the action is "to allow the possibility of hierarchical planning". But for that to be so, I would expect the effects of the REQUEST action to be a subset of the effects of the body of the request action. But I don't see that that's so. Suppose I had maliciously attempted to get you to do something foolish, but had been detected. We would mutually believe that I wanted you to do that thing, but surely that wouldn't count as having requested you to do it?

Allen gives examples of chains of inferences that can be made from the fact that the user made speech acts of some surface type to the fact that he made a speech act of a perhaps different type. To do

this he has to add a new inference rule.

```
Hr b Sp w mb( Hr, Sp, Sp w X )
Hr b Sp b mb( Hr, Sp, (Hr b Sp w X) -> (Hr b Sp w Y) )
-------------------------------------------------------
Hr b Sp w mb( Hr, Sp, Sp w Y )
```

and new heuristics to govern it. These say that deductions about what is mutually believed should be preferred as long as there is only one deduction to make, but if there is more than one possible, none should be made. This is sensible. Once an invisible choice about what to deduce has been made, the result cannot be mutually known, and neither party can have expected to rely on its being mutually known.

Allen's methods were incorporated in a large NLU program built at BBN (Brachman et al, 1979). The only significant differences were these:

- Belief and contexts spaces were implemented in a ways that turned on the features of KL-ONE, the project's knowledge representation language.

- Some words in the input were treated specially to change the system's expectations about the user's goals. For instance, when the user said, talking about the part of the ATN displayed on the VDU, "No, I want to be able to see the S/AUX", the "no" created the expectation that the user wants the display to change. As a result, in the example they give, the system chooses to keep but move what is displayed so as to show S/AUX, rather than to display the node by itself.

- The inferences can be short-circuited. The special case of the user telling the system that he wants something is recognized by the

parser. Where Allen would have had the two-step inference (the inference runs downwards)

Hr b Sp w surface_inform(Sp, Hr, Sp w Q)
|
Hr b Sp w Hr b Sp w Hr b Sp w Q
|
Hr b Sp w Hr b Sp w Hr w Q

the BBN system goes immediately to

Hr b Sp w surface_inform(Sp, Hr, Sp w Q)
|
Hr b Sp w Hr b Sp w Hr w Q

2.5.3. Problems with illocution

Allen offers this relation between "surface" and "illocutionary" speech acts: From the performance of a surface speech act, the hearer deduces a chain of goals. One of these may be the same as the body of an iilocutionary speech act. A speaker who wants the body of an action wants the action. So he also wants the action's effects. Now two things are possible. One can name the illocutionary speech act that has been performed; or one can carry on inferring ulterior goals. But it is not clear that the first is necessary to the second. (Cohen and Levesque (1980) make this point).

Consider the definitions of REQUEST and SURFACE_REQUEST again. The illocutionary REQUEST is there so that one can infer from the mutual belief

mb(Hr, Sp, Sp w Hr do Action)

to a credible goal for the speaker, to wit

Hr want Hr do Action

(Remember this is applied in the context "Hr B Sp w ...", so the hearer is not in the odd situation of deciding what he himself wants). But one doesn't need to apply the body/action plan inference rule to do this. In the simple example of interpreting "Pass the salt", Allen lists the inferences as

```
        Hr b Sp w surface_request( Sp, Hr, pass( Hr, Sp, salt )))
                  |
                  |                              action/effect
                  |
      Hr b Sp w mb( Hr, Sp, Sp w pass( Hr, Sp, salt )
                  |
                  |                              body/action
                  |
      Hr b Sp w mb( Hr, Sp, Sp w pass( Hr, Sp, salt )
```

But the hearer should be able to recognize the speaker's goals from line (2). Allen provides the "decide inference"

```
        Hr b Sp w Hr b Sp w  P
        ----------------------
        Hr b Sp w Hr w  P
```

Why should one not also have its mutual belief analogue

```
        Hr b Sp w mb( Hr, Sp, Sp w P )
        ------------------------------
        Hr b Sp w Hr w P
```

Now the hearer could apply that to line(2) to recognize the speaker's final goal. Then line (3), recognizing the illocutionary act, is otiose.

But if Allen doesn't allow himself this, then he does need illocutionary act recognition if he is to allow what speaker says to have any efect on what hearer does. Presumably hearer will only respond to what he thinks speaker really wants - to X in the the

context "Hr b Sp w ...", as in "Hr b Sp w X". But there are only two ways to get things into this context:

- an inference rule that explicitly changes context, as suggested above.

- by proceeding indirectly. Suppose there is an act Act like this:

preconds: .....

body: C1 X

effects: C2 Y

where C1, C2 are contexts. One way of changing context is by doing these inferences

Hr b Sp w C1 X

| body-action

Hr b Sp w Act

| action-effect

Hr b Sp w C2 X

Then if Act is in fact REQUEST, one can go from a goal of the form

Hr b Sp w mb(Hr, Sp, Sp w Hr do X)

to one like

Hr b Sp w Hr w Hr do X

One still has to get this to the form "Hr w Sp w Z". The details of this will depend on X and Z.

Nevertheless one will want to recognize illocutionary speech acts for other purposes. For instance, if A wants to answer B's question "Did you hear my request?", A has to be able to spot the request to know

what B is talking about.

Cohen & Levesque propose a way of recognizing speech acts that uses the body/action rule, but which does not make this essential to recognizing speaker's intention. For instance, REQUEST(Requestor, Requestee, E) is

effect: (MB y x (WANT y E))

body: (MB y x (WANT x (BEL y (WANT x E))))

prereq: (AND (MB y x (CAN y E))

(MB y x -(WANT y -E))

(MB ẏ x (HELPFUL y x)))

(I trust the notation is obvious).

They say "If the prerequiste holds, any action making the body true achieves the effect". While the general approach seems right, I doubt that an action body is just a proposition. I would guess it would be some set of actions, intuitively a more detailed description of eg REQUEST. If it is not, why can't the action be written as

effect: (MB y x (WANT E))

body:

prereq: (AND (MB y x (WANT x (BEL y (WANT x E))))

(MB y x (CAN y E))

(MB y x -(WANT y -E))

(MB y x (HELPFUL y x)))

The difference between "John murdered Tom" and "John killed Tom" must reside, not in the preconditionss or effects of the act schemata, but in the description of which acts with those preconds and effects count as murders and killings.

2.5.4. Differences between Allen and me

The major difference between what Allen did and what I propose is this:

For him, an utterance works by revealing the speaker's plan to the hearer. This may reveal previously unknown goals of the speaker's. If the hearer is cooperative, the hearer may try and help him with them. So the speaker benefits.

For me, an utterance works by changing a plan, either the hearer's, or what is known of the speaker's. This change may be intrinsically beneficial, either because it altruistically improves the hearer's plan, or because the hearer's plan is selfishly made better for the speaker; or it may be remotely beneficial, since it makes a change in what is known of the speaker's plan. Then comparison of how the speaker's plan was, and is now, thought to be may reveal new beliefs and values that the speaker entertains, which in their turn may have an effect on the hearer. Goal recognition of Allen's type is, for me, the special case in which what the speaker says alters the plan that he is thought to have, so revealing values he was not previously known to hold, and which he hopes the hearer will help him with.

The sort of thing that Allen's account can't even in principle deal with is that where what speaker says affects a plan but doesn't involve revelation of a goal. For example, if A says "It's about to rain" in these two contexts the effect of his utterance is quite different.

A: Shall I get the laundry in?

B: It's about to rain.

A: Shall I water the vegetables?

B: It's about to rain.

The difference must reside in the effect that the statement has on B's obvious plan. But to find out what the effect is requires a notion of how new facts and goals interact with extant plans. That needs knowledge about how plans can be affected, and about the real world.

2.6. Attack on explanation by script-like knowledge

Here is a quote from Schank (1980) where he reviews some of the motivation for his earlier work.

> "We began to focus on the problem of inferencing intention [...]. We got into this problem because of a peculiar use of language that we happened to come across [...] The example was
>
> Q: Do you want a piece of chocolate?
>
> A: I just had an ice cream cone.
>
> Clearly, it is necessary to understand the answer given here as meaning "no". In attempting to figure out how to do this, we realized that it was necessary to fill out the structure of the conceptualizations underlying both sentences so that a match could be made from the answer to the question."

This "attempt to figure out" has become what can fairly be called the Yale school of language understanding. It descends from the work of Schank and his students and colleagues. Its results appear in (for example) (Schank & Abelson 1977) (Schank 1980) (Schank 1982). Central

themes are

- knowledge-based parsing.
- a rich collection of semantic primitives and ways of combining them.
- an understanding of intentional behaviour that relies on "scripts".

It is this last that I want to consider. The Yale school's ideas have been refined and altered for about fifteen years, and the notion of "script" has not stayed the same, nor has it continued to play the same role in explanations. But these things can I think be said about it:

- Agents are masters of parameterized knowledge structures: scripts.

- Understanding experience is largely a matter of recognizing it as an instantiation of one or another script.

- Scripts will be justified by some deeper level of explanation; for instance, if the structure describes agents' actions, there may be a reason for the structure being the shape it is in terms of the agents' intentions; but that is not vital for its use in understanding experience.

Despite the excerpt, the Yale school considers, not discourse, but stories. Instead of such texts as

At cinema entrance.

Boy: A ticket please.

Ticket seller: You're under sixteen.

they consider those such as

A boy asked for a ticket at a cinema entrance.

The ticket seller thought the boy was under sixteen.

The ticket seller refused.

Is it possible to see how in principle techniques of script-based story understanding could be adapted to discourse understanding? I think not. Or rather, only as useful heuristics for reaching a proper explanation. Scripts have no explanatory force of their own.

But has it been claimed that scripts are irreducibly explanatory? After all, Schank points out "Scripts are really just prepackaged sequences of causal chains." (1980:253). So surely a script is explanatory because it was built in accordance with a prior theory of intentions, and so any actions that instantiate that script will also instantiate some theorem of the theory?

Clearly; but he also claims "plan-based processing is different in kind from script-based processing" (1977:99) and that "There is a causal chain [in a sequence of events that is to be understood by script application] but inferring it bit by bit is impossible, which makes scripts necessary." (1980:253). So if explanations of an agent's actions in a context are to be founded on his beliefs about that context, and if certain beliefs are only available to him when he knows about certain scripts, then such explanations make irreducible use of the agent's knowledge of scripts.

Now it seems that (in discourse at least, whatever happens in stories) there are crucial events - types of speech act - that must

be recognizable if script application is to be possible; but they can be recognized

- either in stories but not in discourse
- or by a version of script which has been so weakened that recognizing one is not different from recognizing a straightforward plan.

Scripts will then be either insufficient or unnecessary to explain the remark's function in the exchange.

The central objection is that when one describes what people say, a reported-speech account has to use speech act verbs, but a quoted-speech account does not. Consider an example from (Schank & Riesbeck 1981:150). The story in question is

> John wanted Bills bicycle.
>
> He [...] asked him if he would give it to him
>
> Bill refused.
>
> Then John told Bill he would give him five dollars for it.
>
> [...]

To understand this, the sentences have to be read, converted to the language of semantic primitives, and then matched with the script that the story is currently supposed to instantiate. But see what has happened; the story has made explicit that what John said to Bill was an asking. The translation of "John asked Mary for her book" is (following Schank and Abelson 1977:157)

John MTRANS (ATRANS book) to Mary

Such a structure is then matched with that predicted by the script, and the purpose of the asking can be recovered. That structure can be recovered because the analysing system knows not merely how to

translate an asking, but also that it is dealing with an asking.

But if the system were faced with the discourse, all that it would know about (for instance) the ticket seller's remark "You're under sixteen" is its literal content and its maker. Suppose that the script about cinema ticket selling demands either a refusal (or an acceptance) at this point. How can it tell that it has got one? There is nothing intrinsic to the remark that makes it a refusal. "Your'e under sixteen" can be used in countless ways. If the script demands that the utterance make itself known as a refusal, then it will be inapplicable.

But perhaps scripts can be refined so that they do have some way of recognizing a refusal. Scripts can have embedded scripts. Why should the ticket selling script not include another, the "refusal" script, perhaps as an alternative? A refusal script would perhaps look for an action that made itself known to be an assertion of -F, when F was a precondition of some plan P. P would be a parameter of the refusal script, and would in this case be bound by the embedding ticket selling script.

One way of doing this is to create mini-scripts, where events are characterized in terms merely of how agents' goals and actions interact. Such scripts are content-free; they are independent of particular actions and situations, such as going into restaurant. This has been a popular approach. For instance Dyer's thematic abstraction units (TAU's) (Dyer 1983) or Lehnert's plot units (Lehnert 1982) are the sort of thing I mean.

But then what is the ticket selling script doing? It is unnecessary for saying what the utterance is called. One would only need that to confirm the finding of the expected "refusal" event, in order to

confirm that the ticket seller's actions were still following the ticket selling script. But to recognize the refusal one must have had a depth of understanding of the seller's plan P at least as great as would be provided by grasping the script he was executing. A script mentions only observable actions, not preconditions, which are what must already have been grasped here. Equally, the script is unnecessary to see the effect that the utterance had on the participants' plans. Seeing that was in fact part of the recognition.

I conclude that script application is at most a (perhaps very powerful) heuristic for finding purely plan-based explanations of utterances designed to affect beliefs and values. It has no explanatory force of its own.

On the other hand, there may well be a place for irreducibly script-like explanations of actions which are purely textual - utterances that mark change of topic, or turn taking, or utterances that constitute the producing of the sort of features of text reviewed in (Levinson 1983:6.2). There for instance he cites the way that the length of a pause before laughing at a joke can be crucial to the occurrence of the event of a joke falling flat. Postponing a laugh to flatten a joke is something one only knows how to do if one knows the expected order of events after a joke. Such knowledge is intrinsically script-like. But such actions are somehow ancillary to actions describable as attempts to change their hearer's beliefs and values, even if they are realized in the same physical utterance.

Chapter 3. Rational Action

I claimed that utterances must be seen as actions that are part of plans. Discourse is supposed to cohere because it is a example of rational action. Rational actions cohere because they are parts of rational plans. This chapter goes into more detail about how actions in general compose plans. Once that has been done, it will be possible to go on and explain how plans are sensitive to utterances, and therefore how utterances can be eplained as actions that are intended to affect plans. So this chapter, though it has intrinsic interest as a description of action, is essentially preliminary.

3.1. Rational action

Rational action is a vague concept. It is nevertheless a vital one for explanation of human action, and needs to be sharpened up. How?

Roughly, an action is rational if it part of a set of actions that its performer believes is the best way of getting what he values. This supposes people are able to look ahead at the states of the world that will follow a sequence of actions and choose the ones that make the world end up the way they want it to be.

Now one needs a way of describing such sequences of actions, and criteria applicable to such sequences which licence calling them "rational". The sort of description I shall use is the plan.

The critical features of rational action appear to be
- that it is directed to benefitting its agent - its end is correct
- that it will contribute to benefitting its agent - its means are correct

A plan is a structure that shows how an action included in it satisfies these criteria. The sort of plan that I shall work with is the and-or tree, in which the nodes are the sort of thing that can be true or false and wanted or feared, and where a father node is true if its son nodes are true.

3.2. Goals

Suppose one asks a person "Why did you do that?", and insists on an answer of the from "Because I wanted such-and-such". Then one can ask "And why did you want that?", and demand the same sort of answer. This game can go on for a long while, but not forever. At last he will have to say "Well, I just wanted it". What are the sort of things that a person can "just want" without being able to give a further reason? Are they all the same? Do they have something in common? There have been lots of answers. Some of them are:

> they are morally good: that is, there is a rule which applies to actions or states of affairs (SOAs) which identifies some as good, where part of one's understanding of "being good" is that good things should be goals of actions. The rule may be concordance with the revealed will of God, or being conducive to the greatest happiness of the greatest number, or being the outcome of that which one would will to be a universal law. The source of the rule is immaterial.

> they cause physical pleasures or remove physical needs: warmth, sexual pleasure, the ending of pain or hunger are of this sort. Perhaps it is impossible for the sort of beasts we are not to take these as goals of action, even if they

can be ignored in the face of more important goals.

they are aesthetically/emotionally pleasing: they provoke a sensation of a particular kind which is distinct from that brought about by physical enjoyments, but which when present is undeniable.

> they tend to preserve our, or our fellows', life: perhaps animals that arose as the result of natural selection must necessarily be constructed so that they act for this reason.

I suggest that even if these are all radically distinct, as far as their being springs of action goes, they are identical. They confer goalhood on (say) a SOA. What does this mean? That action is directed towards it. Surely this leads to a circularity - rational action is directed to what rational action is directed to.

Not necessarily. There are two rejoinders possible:

> We have a prior understanding of what goals are. Indeed, since we are goal-directed ourselves, we have an excellent introspective understanding of them. It's then a separate matter to find out what our goals actually are. (Moral enquiry could be conceived of as finding out about goals we have but don't know about, or about goals we will have as soon as we are told about them).

If we accept the idea of rational action, we can construct the idea of value out of this: "valued" will merely mean "being a goal of action". But then this theory-laden term may well be coextensive with terms from other fields: perhaps with those listed above. That this is always so is

an interesting falsifiable hypothesis.

At any rate, I am going to suppose that we can identify things we value. But "being valued" is not the same as "being a goal". A goal only arises when something valued is not the case. I may be glad to own a car, but if I already have one, and I don't value having any more, I won't spend time and effort getting another. At first approximation, a goal is something both valued and false. This however is too simple.

3.3. Fears

Firstly, there are fears as well as wants. I fear toothache or damnation or bankruptcy. But this is not hard to handle. Fears and wants can both be construed in terms of value:

I want X = I value X

I fear X = I value -X

So if the conditions for X being a goal is:

X is a goal = I value X & -X

then of course:

X is a goal = I want X & -X

-X is a goal = I fear X & X

3.4. Abstaining versus Doing

Secondly, just as there can be reasons for doing an action, so there can be reasons for not doing it. For instance, I won't pick up a hot

iron, because it will burn me, and I fear this. Something bad and now false would become true. Similarly, I won't take my purse and drop it down a manhole. Something good and now true would become false.

(In fact these two sorts of reason are the same. The first transition is

- hand burnt --> hand burnt

FEARED is FALSE FEARED is TRUE

and the second:

have money --> - have money

WANTED is TRUE WANTED is FALSE

But since the negation of anything feared is wanted, the first can also be seen as:

- hand burnt --> hand burnt

WANTED is TRUE WANTED is FALSE

just like the second.)

Whether an action is done or not is decided by the merits of the SOA that ensue. The more valued things that are false after the act, the worse; and the more feared thing that are false, the better. These things are REASONS for (or against) an action. A reason for an action is its effect on the truth of something valued.

3.5. Preventing versus Achieving

In fact what I just said is false. Certainly one decides whether to

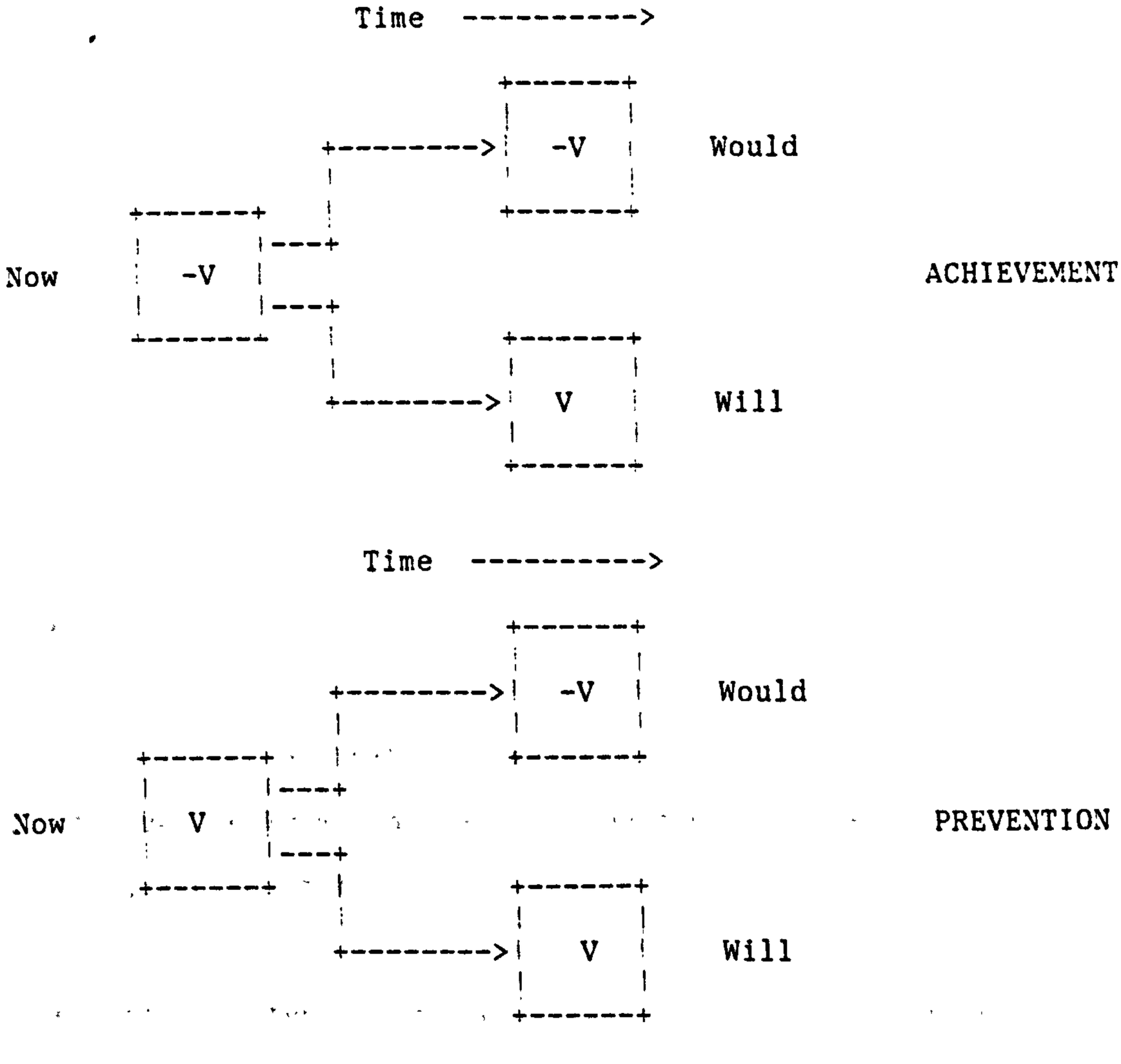
do something or not by comparing the SOA that ensues with some other SOA. But that other SOA is not "the world as it now". It is "the world as it would be if the act isn't done".

For instance, suppose I have a car, and I like having it. If I see a car thief trying to enter it, I'll try and stop him. Why?

Suppose I compare the world before I stop him and after it. Before, my car isn't stolen. After, my car isn't stolen. There's no apparent benefit: indeed, the effort I've put into stopping him may well count against the merit of the "After" SOA.

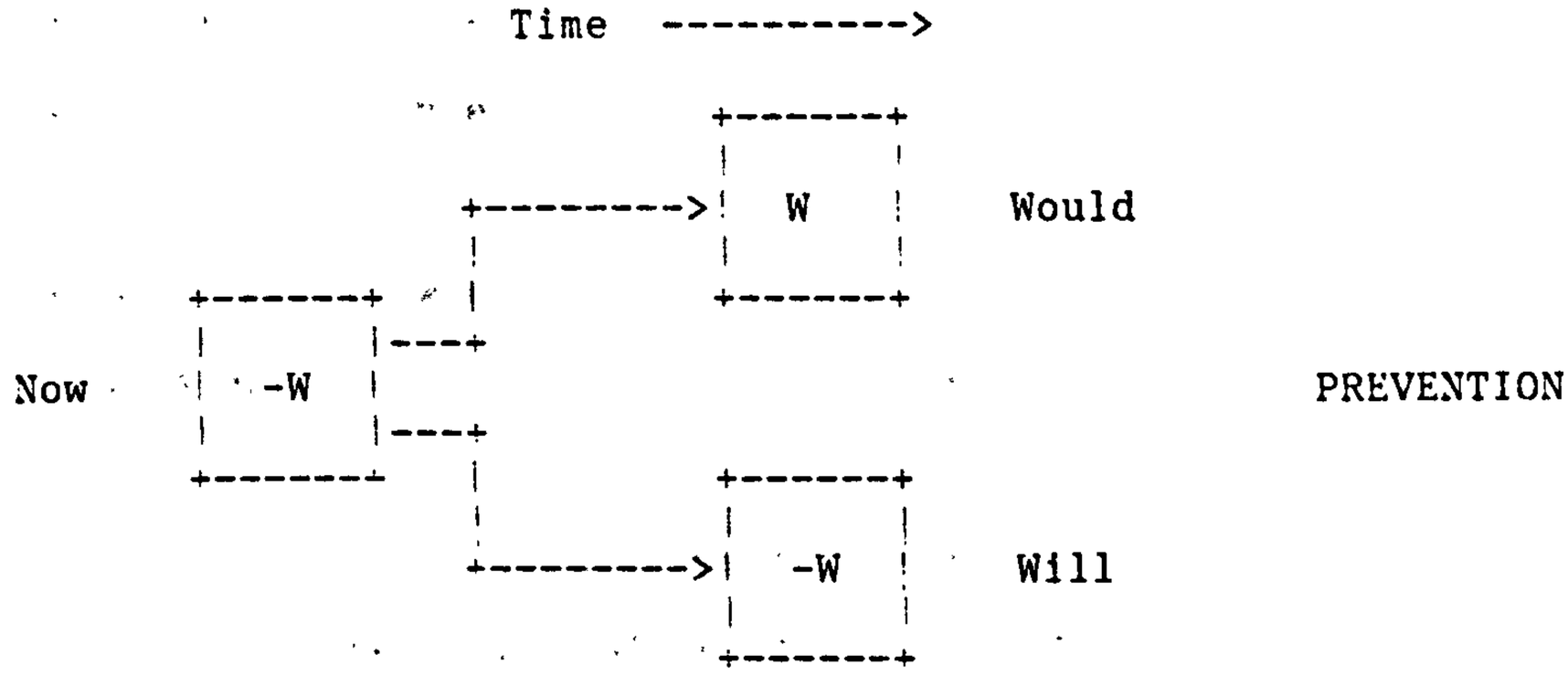
If on the other hand I contrast the world as it WILL be after my action with the world as it WOULD have been if I hadn't acted, (call these the WILL and WOULD SOAs) then there is a clear benefit. In the WILL SOA I have my car, in WOULD I don't.

This gives one a way of distinguishing achievement from prevention. Look at these diagrams of sequences of possible SOAs. V is something valued:



In achievement, something wanted but now false would have stayed false, except that something is being done to make it true.

In prevention, something wanted and now true would have become false, except that something has been done to stop that. (Or equivalently, something feared and now false would have become true, except that something has been done to stop that. Perhaps a diagram of prevention looks more intuitive if we consider the prevent of something we fear, say W. Logically the diagram is the same, with W written for -V.)



3.6. Beliefs

Values are part of rational action. The other part is beliefs. An agent has got to know what is true, both so that he knows what goals are extant, and so that he knows what he can do with hope of success.

I have just confused "belief" and "truth". I did this on purpose. There may be a distinction between things being true and being believed. It does not emerge in considering action. A person believes something if he thinks it true. If he thinks it true he believes it. Of course he may change his beliefs and say that what he used to believe is false. This is just to say that he no longer believes it; he may still be wrong, or have been right the first time and misunderstood the evidence. Even if there was a useful distinction, one would still have to judge a man's actions against what he believed.

There are (at least) two sorts of things that can be believed:

Simple propositions about the world. I call these states of affairs. They can be true or false. The metaphysics of this claim is unimportant. If you prefer, talk about them as those things that sentences in some ideal language describe.

or even those sentences themselves. How they are used in what I say is independent of the ontology you prefer.

Rules about what propositions follow from which (or, equivalently, which SOAs are consequences of which).

These rules represent, not what "actually" follows from what, but what a person believes follows from what. They tell us, given some state of his belief, what else he will believe. They also enable him to direct his actions: by using them he can predict what will follow if he performs, or abstains from, an action.

Because these rules are epistemological rather than "real" they are of several sorts; but they are all licences for belief. If their holder believes the antecedents, he believes the consequent. I write them like this:

```
Consequent <=
        Antecedent-1 & Antecedent-2 & Antecedent-3 ...
```

These are some of their kinds:

- rules about physical events. These are the epistemic counterpart of physical causation:

```
object falls <=
        object massive & object unsupported

object burns <=
        object has burning point T &
        temperature of object is T
```

- rules drawn from lexicography. These are meaning postulates:

agent is brutish <=

agent careless of causing pain &

agent is stupid

agent buys object <=

agent pays something to someone in return for object

- rules drawn from knowledge of social world:

agent1 dislikes agent2 <=

agent2 sincerely insults agent1

3.7. Representing actions

If I ask you "Why is there tea in that pot?" the answer you give need not be teleological. You may say "Because I poured hot water over tealeaves" We both know that doing that produces tea; or to put it another way, if I believe you did that, then I am licensed to believe that there is tea made. This is the same sort of licensing of belief as I just talked about. But there SOAs were connected. Surely pouring hot water isn't a SOA, but an event, an action. Does this mean one needs a different mechanism to to allow one to infer the effects of an action?

I think not. It seems to me that one can draw a parallel between SOAs and actions:

- a SOA is true or was true
- an event is occurring or did occur

It is possible for an action to be a goal. For instance, I may want to go for a walk, and I may perform actions to this end - for instance, putting on wellingtons. It should be possible to represent this as the goal of a plan, but that is only possible if SOAs and actions are the same sort of thing.

One standard way of representing actions is by defining an action schema. A schema consists of:

- an operation
- a list of conditions which must hold before the operation can be performed
- a list of the conditions that are true afterwards

For instance:

```
Make tea:
Postconditions   have infusion-of-tea
                 - water in kettle
Operation        pour hot-water from kettle to teapot
Preconditions    tealeaves in teapot
```

But this sort of representation is not ideal:

- There are actions that don't fit into this mould at all. Eg, going for a walk. While this may have preconditions, the postconditions - a healthy glow perhaps - certainly don't reflect the important parts of what has been done. Furthermore, what is the operation in the "going for a walk" schema? "Walking" can't be it because one can walk without going for a walk.

- The relation between an effect "have tea-the-infusion" and an action "make tea" seems analytic. If one doesn't end up with tea one hasn't made it. There is no operation "make tea" which can be defined independently of its effects. There is no specifiable course of action a person could follow after which one said "He made tea, but it didn't work".

- A plan may involve both rules and representations of actions. For instance, supposing I want to clear a house of mice. I have to deploy both knowledge about possible actions (buying and laying poison) and non-action knowledge - that mice die if they eat poison. The planning process will be complicated if these bits of knowledge come in different formalisms.

Instead I propose a representation based on the idea of formal cause. The formal cause of an event or SOA is those things in virtue of which one can say the event occurred or the SOA holds. For example, if I say "She is beautiful" and you ask "Why?", I can reply "Because she has blues eyes and flaxen hair and is divinely tall". These facts have not caused her beauty in the way that dropping an egg causes it to be smashed. It is just a lexicographic fact that these facts mean that she is beautiful. Written as a rule:

```
X is beautiful <=
        X is female &
        X has long flaxen hair &
        X has blue eyes &
        X is divinely tall &
```

But similarly actions can be unpacked in terms of other actions at a more detailed level which taken together in certain circumstances are

grounds for saying the gross action occurred. The formal cause of an action is just a specification of these actions and circumstances.

What is involved in, say, "I met Celia off a train" ? I have to have gone to the station, Celia must have alighted from the train, and I must have greeted her. I identify actions and events by prefixing them with an up-arrow ^ . So:

```
^ I meet Celia of train <=
        ^ I go to station &
        ^ Celia alights from train &
        ^ I greet Celia
```

What action takes place may depend on what circumstances it is done in. If I apply a match to a pile of wood, paper and sticks in a garden I am lighting a bonfire. If I do the same thing in a grate in a hearth I am lighting a fire. So:

```
^ light bonfire <=
        ^ apply match to a pile of wood, paper and sticks
        pile of wood, paper and sticks is in garden

^ light (ordinary) fire <=
        ^ apply match to a pile of wood, paper and sticks
        pile of wood, paper and sticks is in a grate
```

3.8. Relations between actions

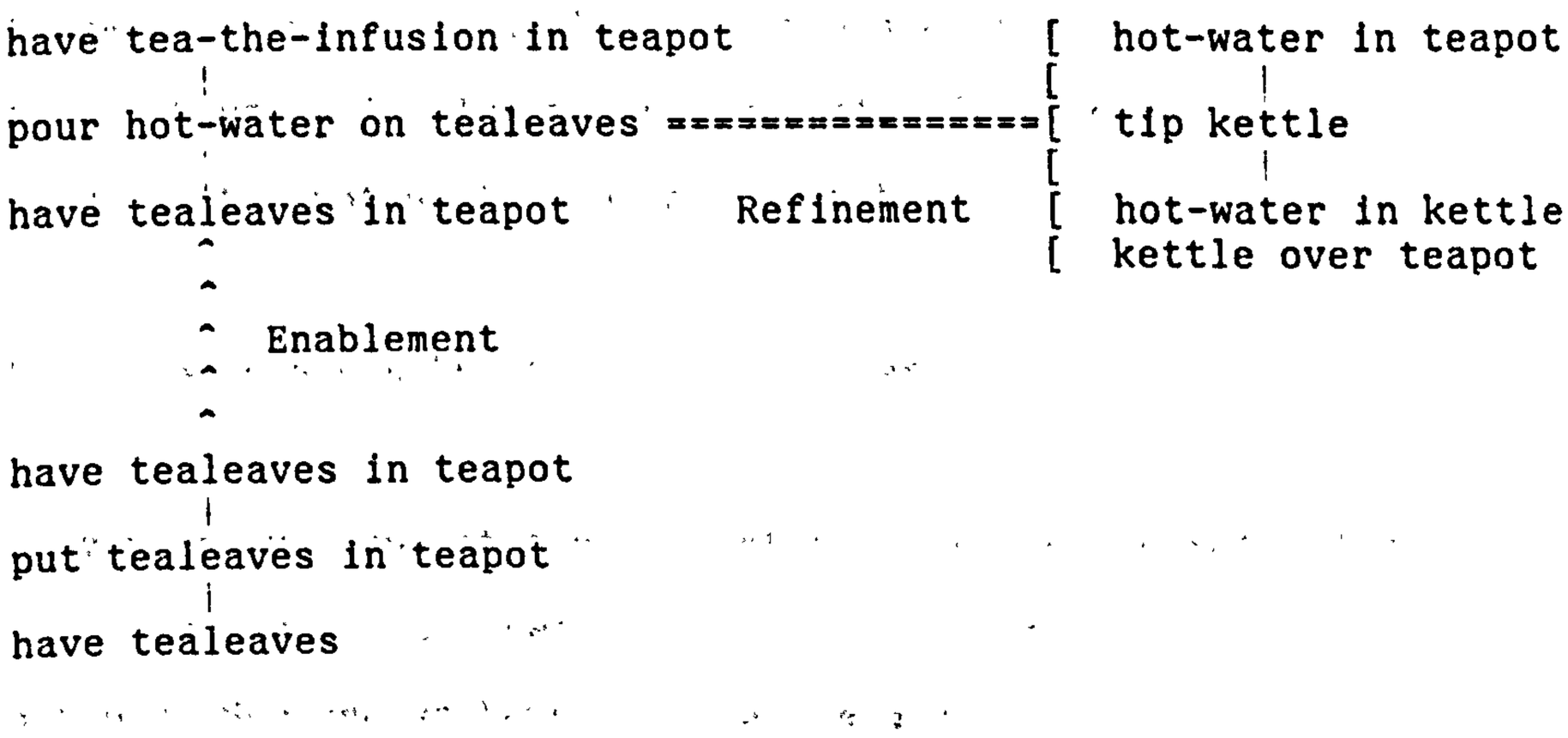
Going back to the schema representation, imagine how schemata will fit together to form a plan. There are two things that could count as

"fitting together", two axes of relationship:

- One schema could have postconditions that were the preconditions of the next. The first is ENABLING the second.

- One schema could be the operation of another. The first is REFINING the second. It is telling one in greater detail how the second is done. This is of course a making explicit of "actions having internal structure"

For example, with what I hope is an obvious notation:



Now how far can these processes be taken? How far in a plan need one imagine enablement and refinement going? Presumably no schema can enable an already true precondition, and so in a plan a chain of enablements stops at preconditions which are either true or incapable of enablement.

How far can refinement of action go? I suspect that the bottom level is something like muscle-fibre motions. Actions are a special case of event in which there is an agent. What the finest-grained sort of events in general are, I don't know. But I don't think it matters. One needn't go all the way down. At some point one can say "This gross action occurred. I am not interested in its internal structure.

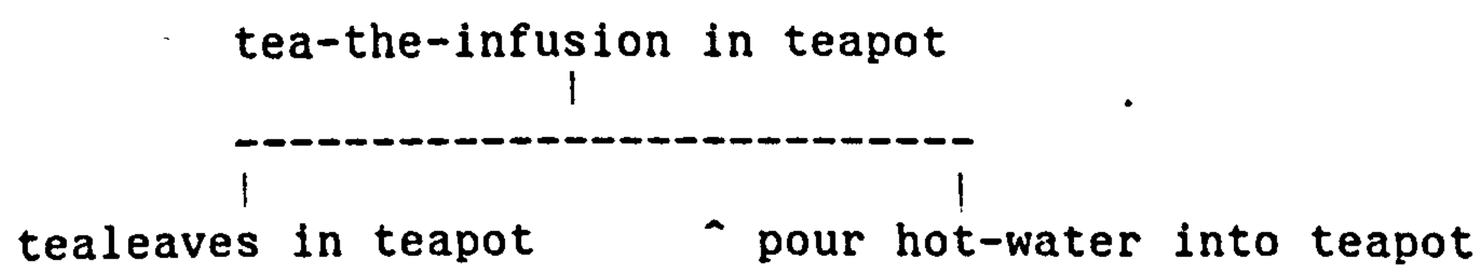
I can make it happen simply by willing it. Here I stop".

Ultimately, one could expand a plan in such detail that its fringe was entirely either SOAs or muscle-fibre motions. At this point one could say both that that the whole plan had genuinely become a contingent proof - if the fibres moved as specified then everything else would happen as expected, and the physical consequences of these movements would show exactly what the effect of each action arose - and that it was credible to see such things as true when wanted.

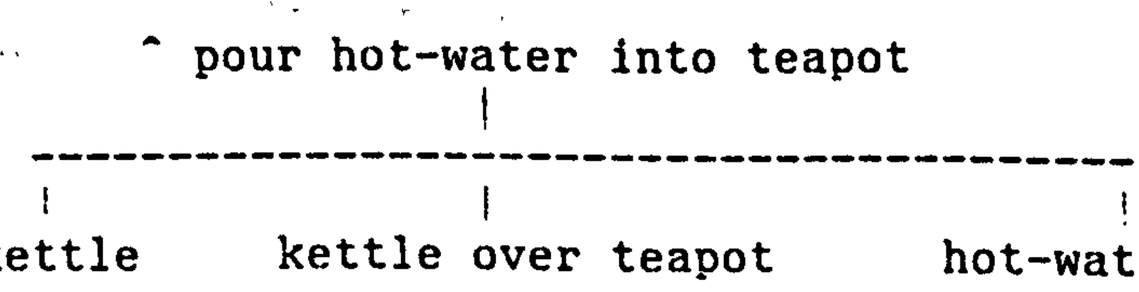
In fact both ends of a plan are usually negligible. At the bottom, there in no point in analyzing actions as far as they will go. At the top, though all value may perhaps stem from some value such as "stay alive", it's not useful to look that high.

3.9. How acts can be said to cause SOAs?

If you are unhappy at the idea of actions causing SOAs, let me try to show how one can get rid of this by looking at the expansion of actions. Suppose we have a bit of an and-tree:



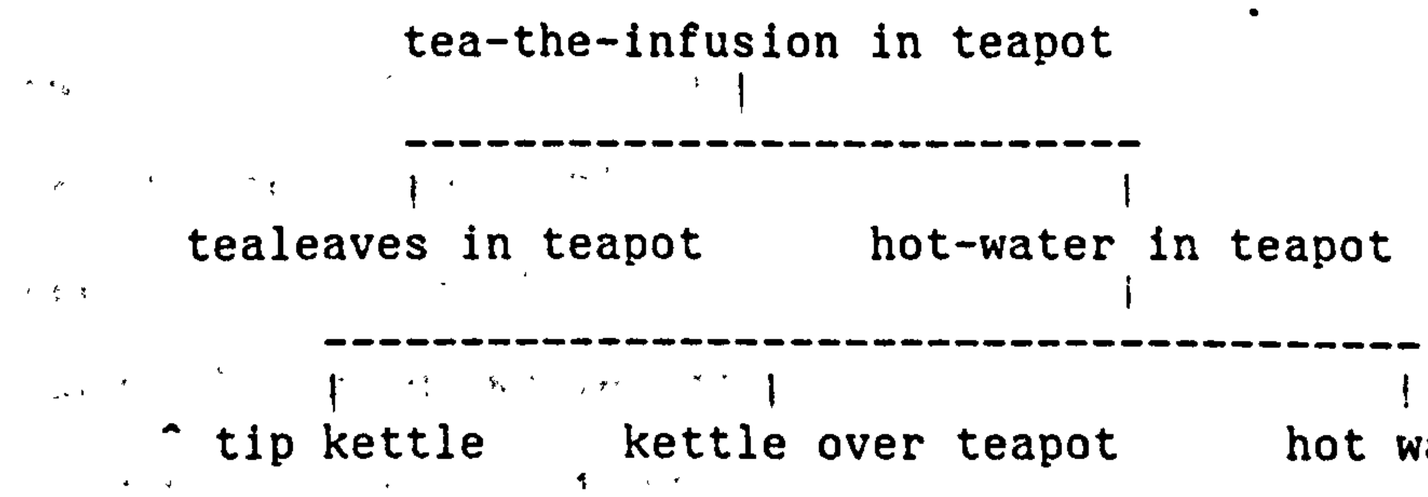
Any gross action can be broken into finer actions and circumstances. So analyse " ^ pour hot-water into teapot":



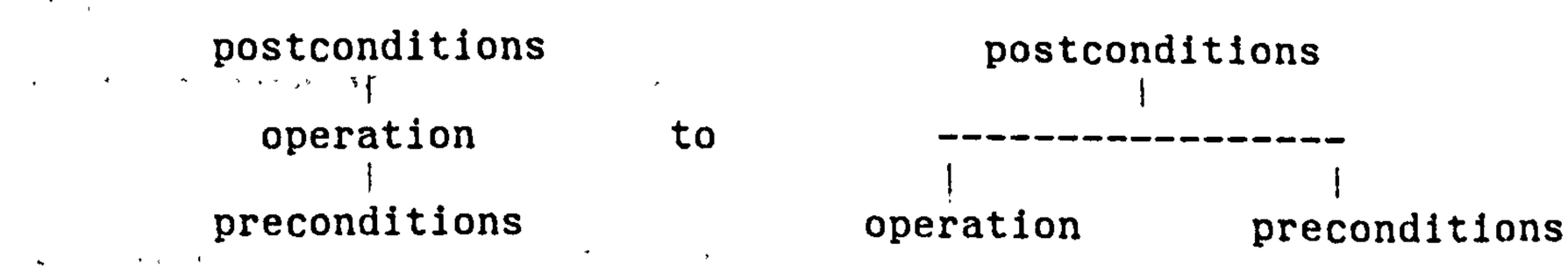
If one accepts that the action " ^ tip kettle" can in its turn be analyzed, then one can allow:

```
hot-water in teapot <=
        ^ tip kettle &
        kettle over teapot &
        hot-water in teapot
```

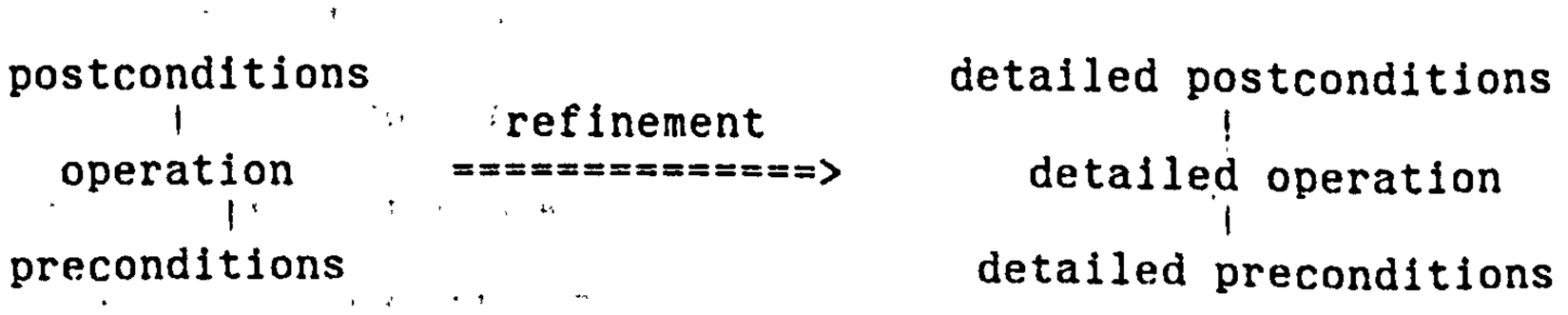
But these antecedents are exactly those that licensed the assertion of "^ pour hot-water into teapot". So if that action was done, then the conditions for hot-water being in the teapot are also fulfilled. If one ignores problems about indicating sequence of events, one could redraw the tree as:



In general there seems to be a translation from schemata-representations to rule-representations like this:

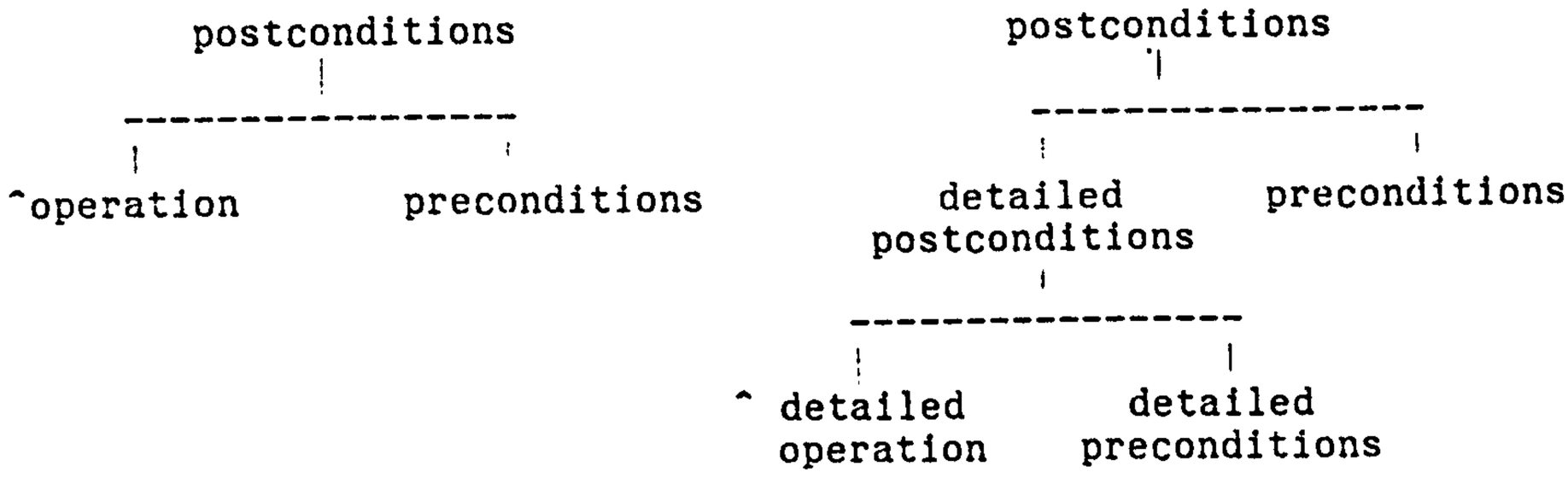


and if the operation is in its turn a schema like this:



then there are two alternative and-tree pictures of a plan using these schemata, which correspond, not to two degrees of checking that the plan's preconditions are enabled, but to showing it at two

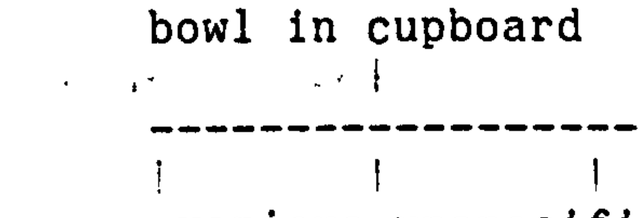
different levels of detail:



3.10. Actions which are defined by their effects

The schema-representation of the structure of an action explicitly includes the effects of the action. In one way the rule-representation can't do this. Consider the action of "putting a bowl away in a cupboard". What are the criteria for an action to count as this? The answer isn't a description of the action's fine structure, because this isn't unique. There are so many ways one can put a bowl away: by hand while gripping it from the side or from below, by sliding it off a tray, by putting away a larger bowl that contains it. There is no feature common to all these sets of actions. What makes them all cases of "putting a bowl away" is the outcome - the bowl's being in the cupboard - and the fact that that events that led to this were actions you performed.

So how can one represent this in rule form? Think back to the case of pouring hot water into a teapot. This action summarized a set of finer actions and SOAs. Those finer parts entailed a SOA (hot water in the pot) that explained the effect of pouring the tea. In the case of putting away the bowl, one knows the entailed state (the bowl's being in the cupboard), but not the fine detail that led to it. One is saying that a proof-tree of this sort exists:



but since the unspecified bits are unspecified, one can say nothing about them. The action "put away" has no fine structure but must be taken as primitive. On the other hand, there is (by definition) a rule:

bowl in cupboard <=

^ put bowl away in cupboard

This is not really sufficient. Imagine me throwing a bowl so that it lands in the cupboard unbroken. One ought to be able to recognize this as a special case of "putting away". Unfortunately to be able to catch this one would need a rule something like:

^ person put bowl away <=

person did an action A &

action A caused the bowl to be in the cupboard

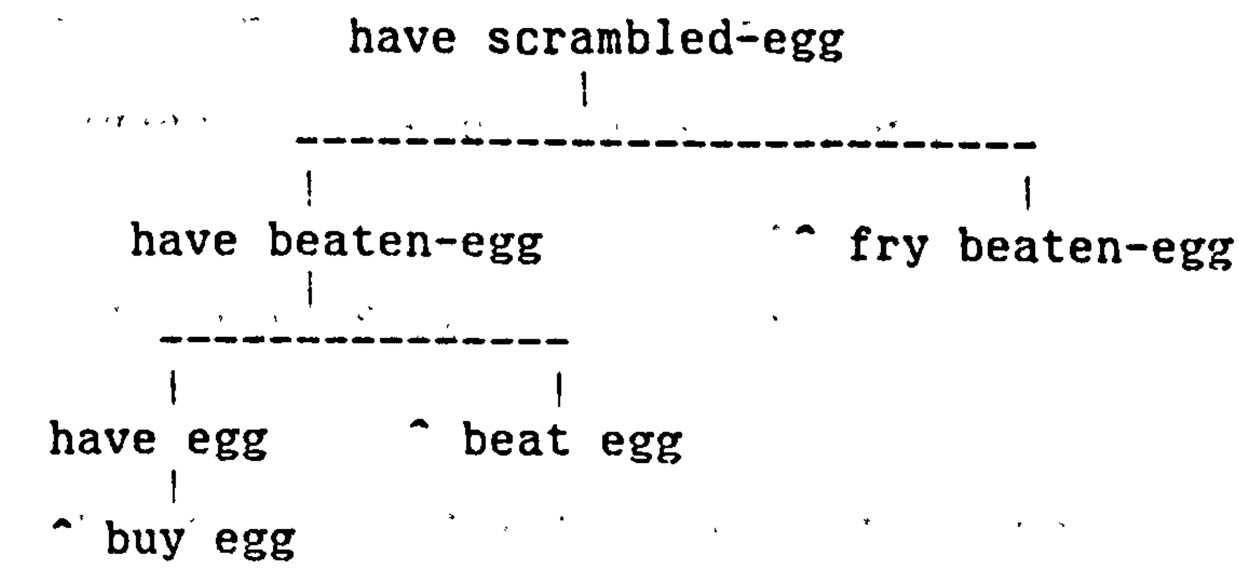
But this involves predicates ("did", "caused") that range over actions. Furthermore, whether A caused B is not something one can assimilate to a SOA or an event; it is essentially a relation between such things. But it could be used if its truth conditions looked at proof-trees. Then one could identify "A caused B" with "A is below B in the proof-tree that contains B". The justification for this would be that if one explicates "A causes B" as "B would not have occurred if A had not occurred" this is paralleled by "B would not be true if A were not true". (I am ignoring the problems about alternative

proofs and causes).

3.11. The relation between proofs and plans

A proof in the form of and-tree is true when all the nodes at its fringe are true. Now imagine an and-tree some of whose fringe isn't true. One isn't licensed to believe in the nodes at the top of the tree. But suppose the false parts of the fringe were made true. Then all the other nodes in the tree would be true too.

This is how I think of a plan. It is a proof with a fringe which includes actions. These actions are not at first believed to have happened. But they have the curious property that they become true if it's desired - or to look at it another way, they can be performed on demand. For instance, if the actions in this tree are performed, it will be a proof tree that shows we can expect to have scrambled egg:



A pure and-tree is too restrictive. One can easily form plans which involve alternatives. One may intend to buy sugar at any of several shops. To allow for this, I treat a plan as an and-or tree.

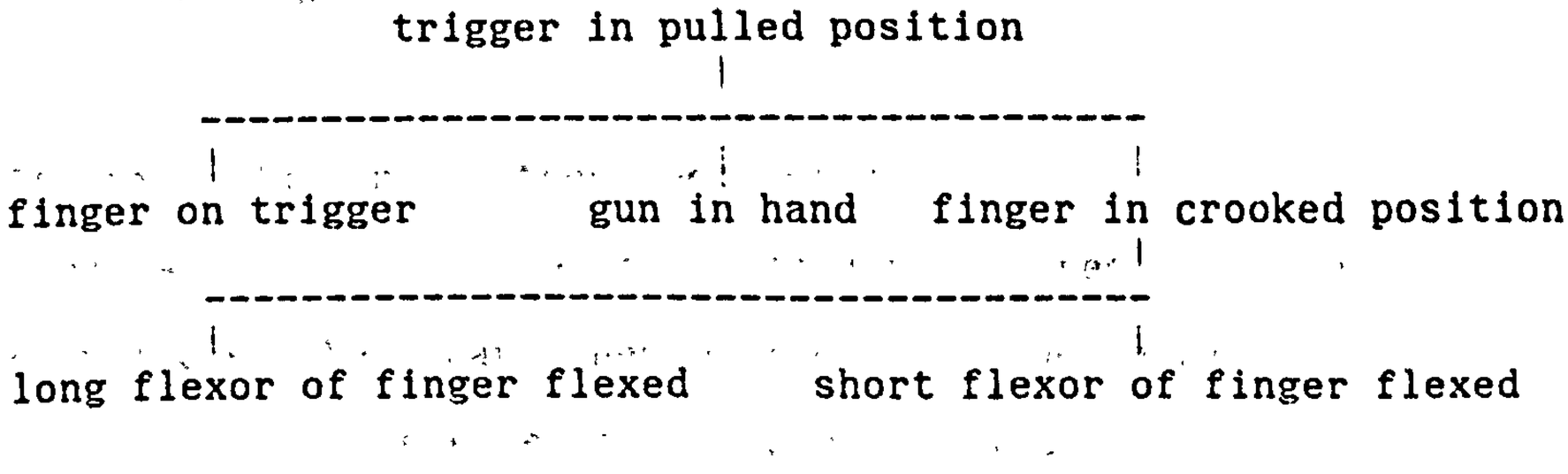
But if a plan has alternatives, one has to distinguish between the tree as a whole, and that subset of it which is actually intended and which will be executed. This subset will reduce to a pure and-tree, and represents what the agent is going to try first. That may of course fail, and at some or-branch another branch will be selected.

It is quite possible that there should be no complete selected plan.

3.12. Why actions behave like facts in proofs

This section attempts to make the the relation between facts and actions more precise, and to explain why both are legitimate nodes in a proof.

Agree that the facts about the states of your muscles are, unlike all other facts in the world, true or false according to what you want. These facts are capable of being the grounds of a proof tree. Their immediate consequences will be facts about the position of your body. These position facts may then have other consequences. Here is a sketchy instance of such a proof:



Now of course these muscle-state facts can become true at some moment when they were previously false. This is what a movement is. If facts can become true, then proof tree can become true. A proof tree is true if the rules that it involves are valid, and all the facts at its fringe are true. Need I be more formal about a proof tree?

I suggest that one can identify an action with the becoming true of a proof tree when that occurs because of the becoming true of some muscle-state facts. What the action is, or rather what you call it, depends on what how you define that action. The definiens of an action will involve predicates over proof-trees and changes in the

truth of proof trees. For instance, one could define "X pulls trigger" as

A proof-tree P became true

P has top node "trigger in pulled position"

All the fringe-node muscle-state facts in P were about X's muscles, and they became true because of X's intention to bring about the top node

Obviously this is not an exhaustive description of the proof tree. That must be right. There are many ways a trigger might be pulled (with teeth, string, toes ...) and the proof trees that correspond to each of them will differ, but all should qualify as trigger-pullings. Under this definition they will.

If this is true, then believing that an action has occurred will involve you in believing that a proof tree has become true. Other deductions that can then be made from the things proved in the new proof tree. These deductions can tell you either about new facts or about different actions.

3.12.1. How can an action support a fact?

This is the case that corresponds to enablement between rule schemata, though the analogy is not perfect.

Why is this rule, where an action supports a fact, valid?

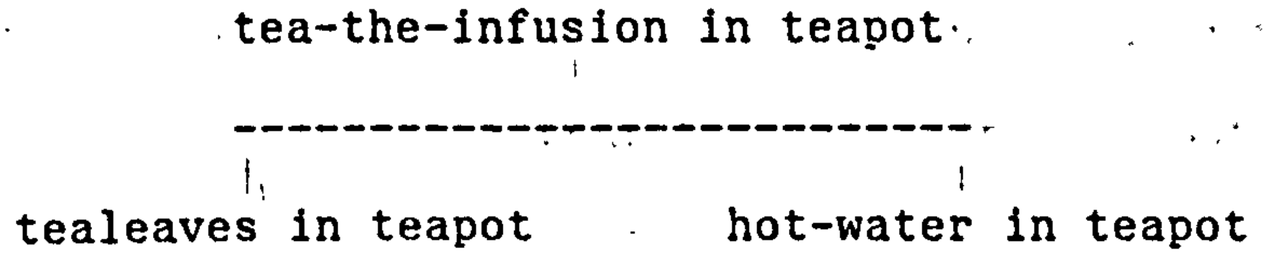
tea-the-infusion in teapot

---

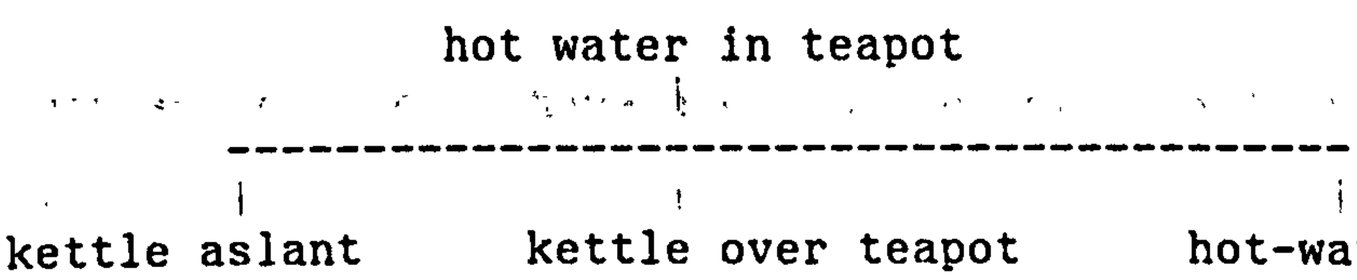
tealeaves in teapot ^X pours hot-water into teapot

First, accept these uncontentious rules, 1 and 2, which contain only facts:

rule 1



rule 2

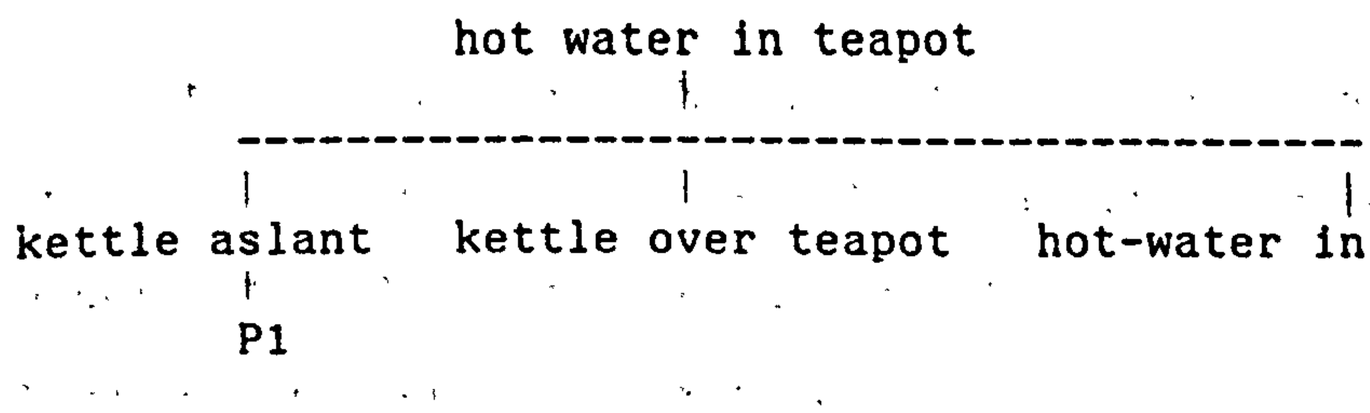


Now, suppose that I define "^X poured hot-water into teapot " as

A proof-tree P became true

P has top node "hot water in teapot"

P looks like this:



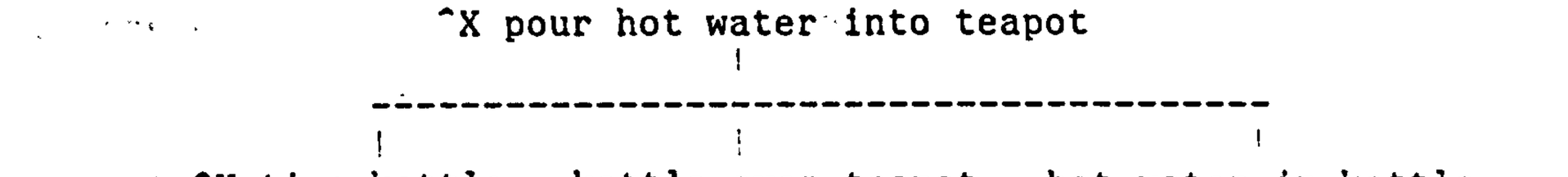
(All the muscle-state facts in P1, a sub-tree, were about X's muscles.)

If I do, then to accept that the action "^X pours hot water into teapot" occurred is to accept that a proof tree like the one above has become true . If I accept this, then I must accept "kettle over teapot" and "hot-water in kettle" are true. If I accept these and the uncontentious rule 1 about tea-the-infusion then I know that tea-the-infusion is in the teapot.

3.12.2. How can an action support another action?

This is the case that corresponds to the refinement relation between rule schemata.

Why is this rule, where an action supports an action, valid?



Suppose we already have a definition of " ^X tipped the kettle " like this:

- A proof tree Q became true

- ... and here follow other conditions on Q. All that matters for now is that they require that Q contains or entails "kettle is aslant"

Then if I accept that "^X tipped the kettle", I accept Q became true, and so I also accept "kettle is aslant" became true. If I further assume "kettle over teapot" and "hot-water in kettle" then a proof tree whose top rule is rule 2 becomes true. But this becoming-true satisfies the definition for "^X poured hot water into teapot". So, when I discharge my assumptions, I must accept the rule in question.

One may want to insist that for X to haved poured water into the teapot, then the water's being in the teapot must be a consequence of X's tipping the kettle. One wants to rule out the case where, while X is pouring hot water from a kettle onto the floor, someone else moves the teapot under the stream. If so, one just has to require that "kettle over teapot" and "hot water in kettle" were true just before and after the moment at which the pouring into the teapot is defined to have occurred. Then the only change that could have made rule 2

fire would be a change that made "kettle aslant" true - for instance the becoming true of Q.

3.13. Action and facts in general

In general, one can relate actions and facts like this:

if ^A is defined as the becoming true of a proof tree P
and P contains the node G
and G,H entail J
then ^A, H entail J

and one can relate action and actions like this:

if ^A is defined as the becoming true of a proof tree P
and P contains the node G
and ^B is defined as the becoming true of a proof tree Q
and ^B contains the node G at its fringe
and the rest of the fringe of Q is H
then ^A, H entail ^B

3.14. Actions defined by intentions

This sort of definition of actions doesn't refer to the intentions of the people who make the muscle movements. But of course there are action words whose definition essentially involves intention. The contrast between "X killed Y" and "X murdered Y" is an obvious example. The account above can't deal with such distinctions. But typically even such an intentionally defined action has an objective component. For instance, the case of Lamb, where Lamb put an empty chamber of the barrel of a gun against the hammer, showed it to his friend, pointed the gun at his friend and pulled the trigger. The gun

went of and killed Lamb's friend. A gun changes which chamber lies under the hammer before firing, not afterwards. On appeal Lamb was held not guilty of murder. The court accepted that X did kill Y by pulling the trigger of a gun, an objective action, definable in the way I outline above, before it went on to consider the question of murder.

I shall say more about intentionally-defined action later.

3.15. What makes a good plan?

Obviously not all and-or trees are plans, and not all plans are good plans. What makes a good plan? Putting it most generally, a plan should be

- beneficial - the SOA it brings about should be better than the one that would occur anyhow. This means a plan should be needed and safe.

- effective - it should in fact bring about the SOA it's intended to. A plan should be sound.

3.15.1. Need

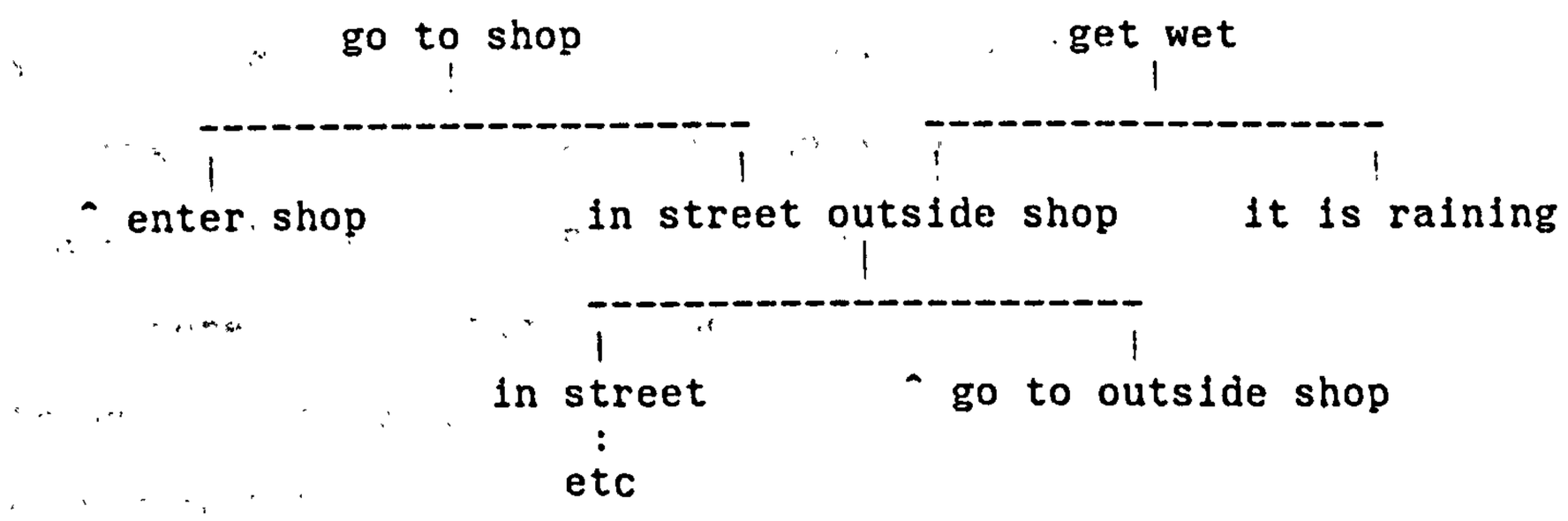
The SOA at the top of a proof-tree will be the plan's main goal. Besides being valued this must also be false, or else it will be part of "what would happen anyhow". This is the requirement that a plan be NEEDED.

3.15.2. Safety

The execution of a plan won't have only one effect. Some of the effects may be good, some bad. The net value of a plan will depend on

the difference between these, so when making a plan one has to pay attention not only to the effect that you intend a plan to have, but all its side-effects as well.

A side-effect is a valued consequence of a proof-tree that grows out of the the side of a plan. An example is the best way to explain this:



Here the effect actually sought is to go to the shop. If all the actions in the fringe of the tree are performed, that is proved. But so also is the side-tree leading to "get wet". If the tree proves the one, it also proves the other.

To determine whether the plan is actually worth performing, one would have to balance the merits and demerits of all the effects. To do this properly one would have to remember that not all values are of the same importance. The merit of getting to the shop may outweigh the annoyance of getting wet. But what I am interested in doesn't need this refinement. One can assume that what one wants is plans without side-effects, or at least, without bad side-effects. A more elaborate model, that considers the net benefit of a plan, will be a strict extension of a model that assumes there is only one effect - nothing will need to be retracted. The requirement that there be no bad side-effects is the requirement that a plan be SAFE.

3.15.3. Soundness

Any plan should work; if one performs the actions it involves, the goal should follow. The most basic condition for this is that the preconditions in the fringe of the tree should be true (if SOAs) or possible (if actions). These two properties seem to be analogous. Just as one believes that some facts are true, so one can believe some actions are possible. I now believe that I can clench my hand, though events might make me abandon this belief - for instance if I am holding something. Such beliefs will probably only be about actions that are immediately under my control. Whether I can, say, travel to London is going to be a matter of proving "^ get to London" from more basic facts. The requirement that all the nodes in the fringe of a plan should be trusted in this way is that the plan should be SOUND.

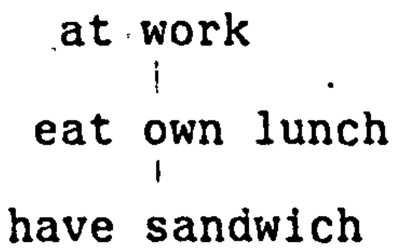
Of course that isn't all that one need to be sure a plan will work. Proof-trees such as I am using are only partial orderings of actions, and if a single agent is going to execute the plan, he must convert this into a linear order. Doing this can be difficult. However, I am going to ignore this problem entirely.

3.16. Flow of truth and value

When drawn like this, one can think of value flowing down the tree from the top, and bestowing value on all its children, and truth flowing up from the children to bestow truth on their parent. The very top is supposed to be intrinsically valued. Subgoals are valued because they tend to produce the main goal.

But imagine a man who is going to work and asks his wife to make him sandwiches for lunch. Obviously his desire for sandwiches comes from the fact that he is going to work. How does this appear in a plan

such as I've described? Since value flows down a plan, presumably like this:



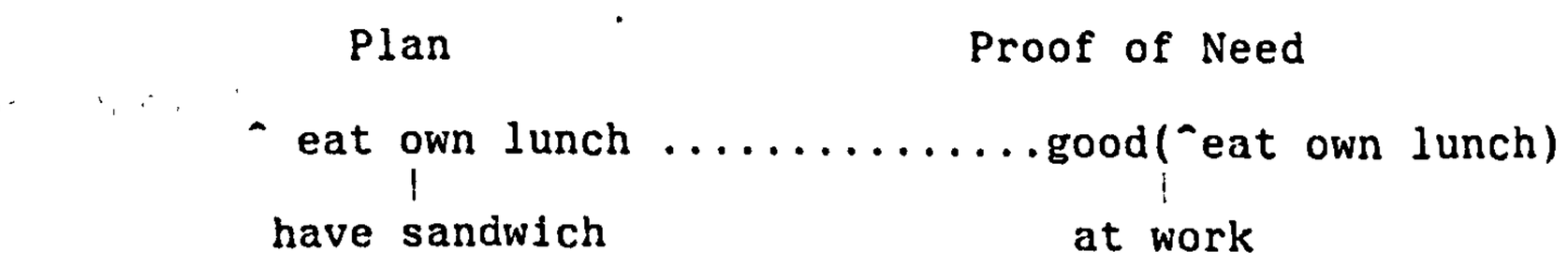
But this is absurd. It implies that eating lunch entails being at work. What one wants to demonstrate is the dependence of a goal on a fact, without always supposing that satisfying the goal entails the fact.

My first attempt at this was to suppose that to demonstrate a plan was rational one should show both

- that there were actions that would achieve one's goal (the and-or tree)

- that the goal was worth getting

If the plan was to achieve X, then there was to be separate proof of "good(X)". Then the plan would have two parts, like this:



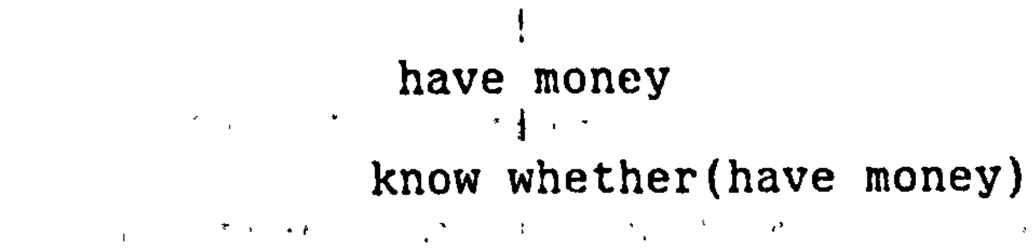
The rule involved in the proof of need

good(^eat own lunch) <= at work

would either be a primitive reflecting common observation or else a summary of a proof that being at work prevented one from eating any other sort of lunch.

However there is another case of goals arising in a way that isn't handled by downward flow of value.

Briefly, if one has a plan that involves, say, having money, one is also going to have a goal of knowing whether one has money. If one doesn't know, one may well form another independent plan to find out. But how does one describe the relation between the initial goal and the knowledge goal? A bit of plan like:



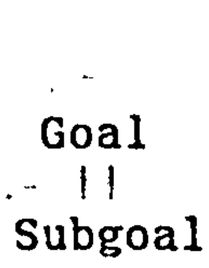
is absurd. It supposes that knowing whether one has money entails one's having money. To avoid this, I imagined a link in plans that says

SubGoal is wanted because Goal is wanted

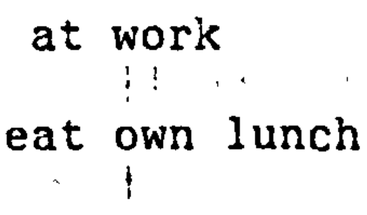
which, so to speak, permits the flow of value down but not the flow of truth up. Or more formally, it is a rule

good(SubGoal) <= good(Goal)

which is a perfectly respectable ordinary implication. I draw it like this



Then the problematic plans above look like:



have sandwich

have money

know whether(have money)

The second appears preferable. It appears to allow one to indicate the flow of value in both cases, while I can see no way of applying the first representation to cases where a need for knowledge arises.

3.17. Plans to prevent and maintain

So far I've talked about plans to achieve SOAs not currently true. But there are plans whose purpose is to preserve something now true which is threatened, or to prevent some bad thing which is now true from coming about. Do these require a different sort of treatment?

No. Preserving and preventing are the same sort of action, since "preserving G" is the same as "preventing -G". They can both be likened to achieving if one asks how one knows when there is something that one needs to prevent. What for instance makes me feel I must prevent myself being locked out when I see Tom about to close the front door? It is a proof-tree like this:

I am locked out

I am outside door -I have doorkey ^ tom locks door

If being locked out is something I fear, the value from this fear is going to "flow down" the tree just as desirability does. If I desire a consequence, I desire the conjunction of its antecedents. If I fear a the consequent I fear the conjunction of its antecedents. I seek to negate what I fear. So I seek to negate the conjunction of the antecedents. This happens if I negate any of then individually, which I do by achieving its negation. So in this case I am going to try and

achieve either "- I am outside door" or "I have doorkey" or "- ^ tom locks door".

If you prefer it done slightly more formally, one can catch the flowing down of value as

```
        X => Y  and  good(Y)  entails  good(X)
and     X => Y  and  good(-Y)  entails  good(-X)
(or     X => Y  and  bad(Y)  entails  bad(X)    )

so if   A & B & C => D
and     good(-D)
then    good( -(A & B & C))
Since   -P => -(P & Q)
then    good(-A) & good(-B) & good(-C)
```

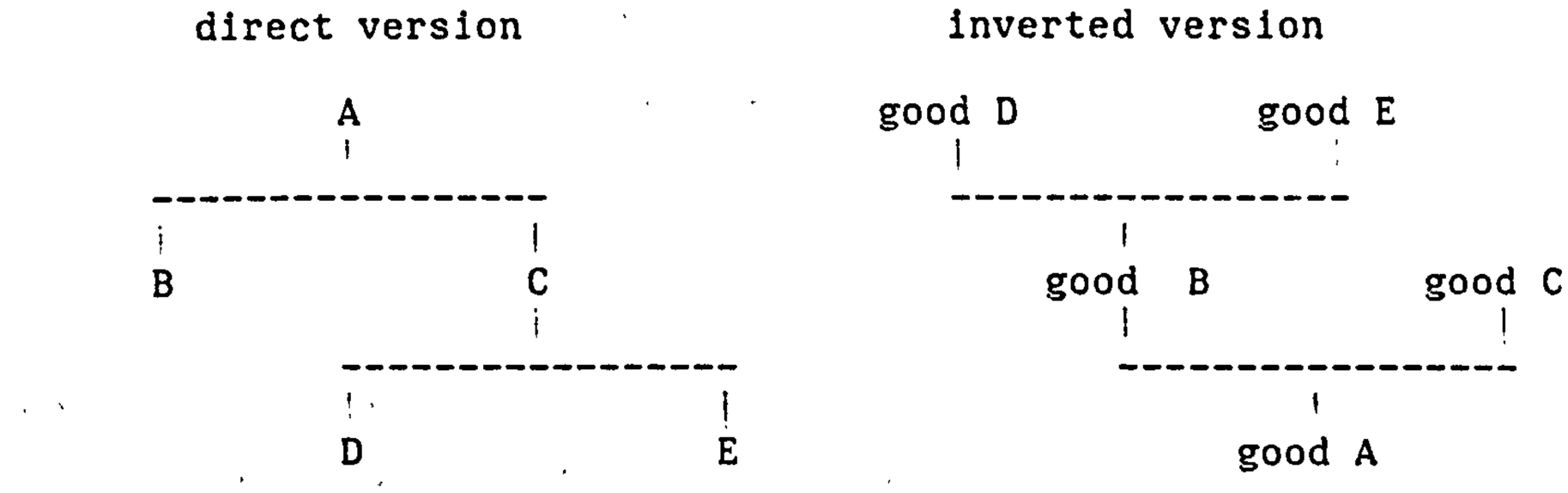
To summarize all this: if you have a fear that something bad will happen, based on a proof-tree leading to that bad, then the negation of any of the nodes in the fringe of that proof-tree may be a goal.

3.18. An alternative method: inverted plans

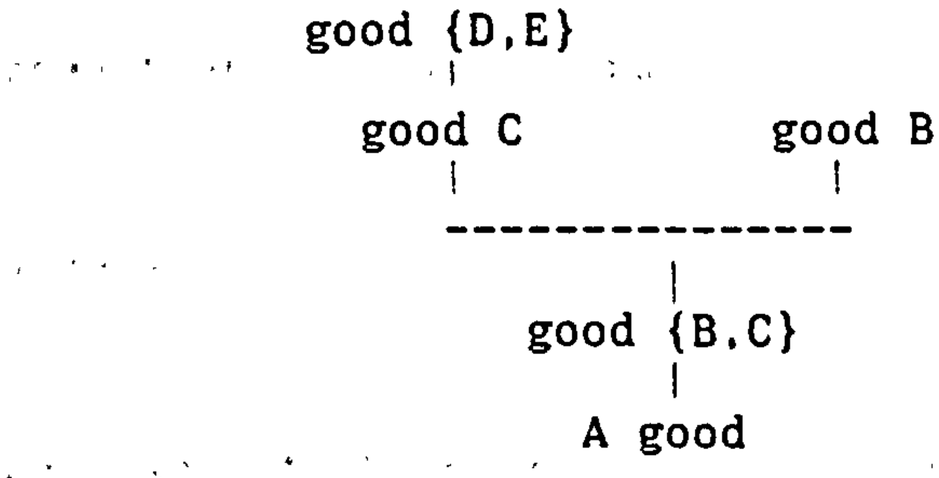
The trick above relied on considering implications between modal statements about the values of SOAs. In trying to find a teleological explanation of an action Action, one is really trying to prove that it was worthwhile. In effect one is trying to prove "good(Action)".

A proof-tree that reflects this is going to be the opposite way up to the ordinary kind. What were low subgoals, which derived their value via every node above them are now high-up conclusions which derive their truth from all below them. Thus (with an obvious notation for

multiple conclusions):



In fact, this is wrong. This suggests that it's worth doing D as long as one does B. But suppose one were unable to do C. D would then be pointless. It is only really worth doing D if one is going to do both B and C. But if we take the antecedents not individually but a set at a time, it looks better:



One could imagine the steps in this proof licensed by a general deduction rule like this:

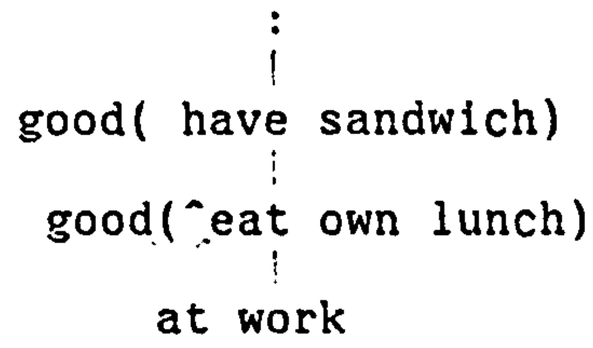
B <= A1 &...& An

good(B)

---

good({A1...An})

One advantage of this approach is that it doesn't need the extra type of link that transmits value only. The bottom of the plan to take a sandwich lunch now looks like:



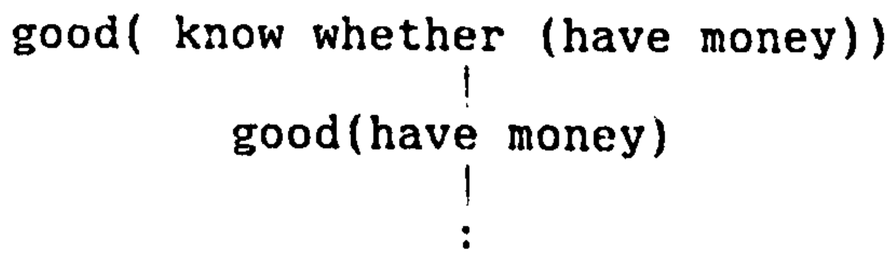
and could continue up to show the goodness of actions that tend to one's having a sandwich. The catch is that now one has no way of telling how the truth of the non-modal part of one node (the part inside the "good") depends on the truth of others. The obvious fix is to let truth flow down. If

good(X) <= good(Y)

then there was also a rule

Y <= X

and so if X is true so is Y. But this runs into almost exactly the same problem as before. In this case



it will make "know whether(have money)" entail "have money".

I prefer the direct version, in which the proof trees are proofs of the truth of sentences, not of their value: not least because it is the more conventional and better understood approach.

3.19. Conclusion

So it is possible to represent plans and proofs alike. There need be

no deep difference between reasoning about states and about events, as long as one ignores any serious representation of time. That can be done as long as the actions in the plan do not interfere with each other. But the picture so far is purely static. Utterances interact with plans and change them. They will do this in the way that new or lost assumptions may affect the validity of a conventional proof. The next chapter exploits the likeness of plans and proofs in describing how such interaction occurs.

Chapter 4. The benefits of changing plans.

4.1. Introduction

What does what I have said about planning have to do with language use? The answer comes in two parts.

- It may only be possible to give the right description of what an utterance does to its hearer by seeing how it affects his plans.

- It may only be possible to give the right description of why a speaker attempted to make the change that he did by seeing that change as an action in a plan of the speaker's.

I'll start by giving some examples of utterances whose whole point is their effect on their hearer's plans. Then I'll show that making these changes can be seen as the speaker attempting a particular goal: to advise his hearer.

After that I shall say more about the ways a speaker can affect plans; then more about the sort of changes he can make; then more about the benefits he gets from the changes.

4.2. Some examples

Let me start with some sample exchanges. They are all based on real examples, though they are simplified.

1 A: I'm going to wash up.

B: I've only just turned on the hot water.

2 A: I'm going to make some tea.

B: I used the last of it this morning.

3 A: I'll park behind that red car.

B: There are double yellow lines there.

4 A: I'm going to have some of that crumble.

B: I wanted to give it to Elaine and Richard.

5 A: I'm going to the library.

B: The electrician is coming at ten o'clock.

6 A: I'll bring you some cocoa.

B: I don't like milk drinks.

7 A: I'll get some sellotape when I'm out.

B: There's some in the cupboard.

What do these have in common? They all start with A saying something about his intentions. Then B says something to make him give up, or at least doubt the usefulness of, his intention. How does B do this?

I suggested that there were 3 things that a plan should be: sound. safe, and needed. What B does is say things, which if accepted, must make A feel that his plan fails one of these criteria. Let me sketch the deduction that A has to make to see this.

1 A: I'm going to wash up.

B: I've only just turned on the hot water.

A plans to wash up. Washing up can only be done effectively in hot

water. A and B know that in their flat the water is only hot if the boiler has been turned on for twenty minutes. B says that he has "just" turned it on. This presupposes it was not on before. So there is not hot water. A's plan has been claimed to be unsound.

2 A: I'm going to make some tea.

B: I used the last of it this morning.

Making tea-the-infusion requires tea-the-leaves. B says he used the last of the tea, which entails that none is left. Again, A's plan is unsound.

3 A: I'll park behind that red car.

B: There are double yellow lines there.

What B says does not at all affect the possibility of what A intends. A can still hope to park successfully. But if he does, then a fact that is a consequence of his action, and the fact B has stated, taken together, entail that A is committing a parking offence and may be fined. A and B share the assumption that fines are bad. B's remark points out that A's action is not safe.

4 A: I'm going to have some of that crumble.

B: I wanted to give it to Elaine and Richard.

Here again B points out that A's action is unsafe. But it does so rather differently. It may be that what B says has no effect on A's expectation of the SOAs that will follow his eating the crumble - mainly that it won't be there any longer. He may have known that Elaine and Richard were coming. But what has changed is that B has put a different colour on these effects by saying that he wants to do something that A's action will make impossible. Things that prevent

the obtaining of goods are bad. B has shown A's plan is unsafe, by stating, not a fact, but a value.

5 A: I'm going to the library.
B: The electrician is coming at ten o'clock.

Again, B states a fact that makes A's action unsafe. A can get to te library. But if he does, the plan of letting the electrician into the house in order to do some repair will be unsound. B announces the existence of this plan by referring to one of its preconditions. (How this happens I'll discuss later.) Again, things that prevent the obtaining of a good are bad. B has shown A's plan is unsafe by stating a fact.

6 A: I'll bring you some cocoa.
B: I don't like milk drinks.

After B's remark A's plan stays both possible and safe. But it become pointless. A will have the cocoa to drink, a fact at the moment false. But as B doesn't want it, it doesn't count as a goal. B has shown A's plan is needless by stating (or rather denying) a value.

7 A: I'll get some sellotape when I'm out.
B: There's some in the cupboard.

A intends to bring some sellotape into the house, in order to do something else with it. B points out that A's goal (strictly A's subgoal, since he wants the sellotape in the house for a further purpose) is not really a goal, since it is true already. B has shown A's plan is needless by stating a fact.

4.3. What good do utterances do?

In those examples, B's intercession in A's plan is rational language use. But why rational? What goal does it serve?

One can explain why people choose to act as they do by pointing to the expectations they have about what is going to happen. These expectations are sequences of SOAs that will or may follow each other if various things occur or are done. People act so as to end up in the SOA they most like. But utterances can change expectations and the values put on SOAs. With new expectations people may choose differently, and expect to reach different SOAs. If Sp prefers these new SOAs, his utterance is explained.

(This assumes that Sp's goals are all SOAs. In fact aims such as being seen to be helpful, which involve the evaluation of actions rather than of the results of actions may be even more important. I shall come back to this.)

From this account, I shall try to extract a description of a process that Hr can apply to see what effect Sp intends. But though this shorter process will able to find that effect, it will not be able to explain why it is intended.

Why cannot the longer, justifying, process be used? Because the formulation that justifies the shorter process assumes that people are always contemplating all possible actions and events. No such process could explain people or animate programs. There are all sort of contingencies that are possible but which they never have in mind. If Sp shouts "Look out, the floor's going to give way", Hr will move off it. One can point to the disaster that Hr acted to prevent, and reify it as a certain sort of proof tree. One can explain Hr's changed choice in terms of his first thinking this proof tree

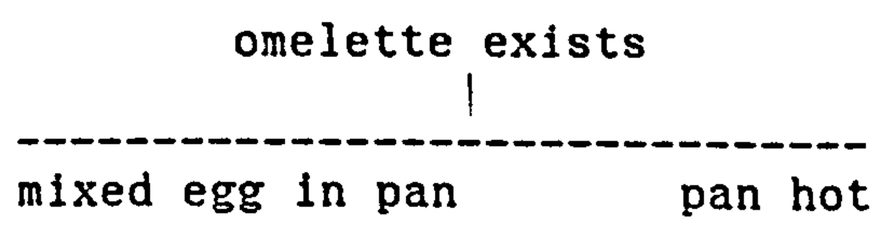
unsound, but later, sound. This odd approach is convenient when it comes to describing the shorter practical process. But it supposes that Hr thought about the proof tree before Sp'S utterance, which is absurd.

The long process comes in four stages. First I will give an abstract and impractical account of how plans are expected to change the world. Then how these expectations lead an agent to make his choice about what to do. Then how an utterance can lead an agent to change his choice. And finally why changing an agent's choice can be good for a speaker.

4.3.1. How execution of a plan changes a state of affairs

A plan can be seen as a complex description of an action. An action is a simple name for a plan. I shall mix my terms.

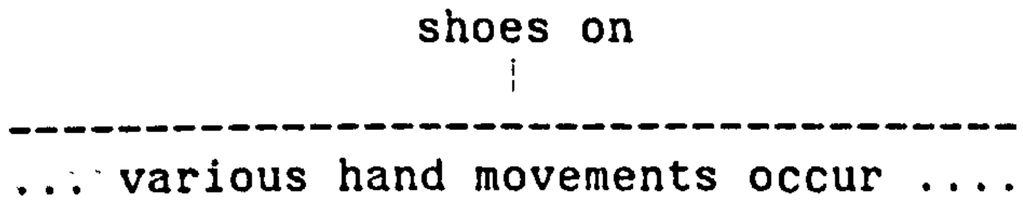
I identified an action with the becoming true of a proof of a particular sort. For instance, to take a joke example, one could identify "fry an omelette" with the coming true of a proof tree like this.



Then frying an omelette in the current SOA (call it NOW) would mean that that SOA had to be updated with the "add-list" of new facts {omelette exists, mixed egg in pan, pan hot}. If the proof tree is true, all the nodes in it must be true, and they will all be found in the updated world.

Because this formulation doesn't include a "delete list" there is a risk of an update producing an inconsistent SOA. It arises this way.

Suppose that "putting on my shoes" is the coming true of a proof tree like this:



But when I put my shoes on, I start with them off. So the initial NOW SOA must contain "-shoes on". If I just add the nodes of the proof tree to the initial SOA, the new SOA will be inconsistent.

In that case the inconsistency was between a fact in the NOW SOA and a node in the proof tree. One could also have a contradiction between a fact in the NOW SOA and one entailed by the proof tree though not in it. For instance, suppose it is raining but I am indoors. My coat is not wet. I go outside. The proof tree of going outside will not include a node "my coat is wet" but this fact will be entailed by the tree. I must be careful to avoid this contradicting the initial fact that my coat was not wet.

The solution is to remember that an update is not a logical assertion. It takes place in time. Suppose that the NOW SOA contains A, and that the proof tree contains -A. Just as -A springs into existence, A goes away. They never co-exist. (Of course in fact actions aren't instantaneous. I shall just ignore the problems that follow from this.)

Here is a definition of the updating process that produces the SOA that obtains after an action. It can fail. If it does, then I will assert that the initial SOA, the rules or the action that lead to this cannot be a proper description of reality. (This is not tautological. What I am doing is preferring to keep my method of updating rather than any particular description of the world. But if I did decide that some set of facts and rules actually described the

world, and the update of some event then turned out to be undefined, then I would have to abandon the method).

Suppose there is a set RULES of rules of the form C <= A1 & ... & An.

Suppose all SOAs are consistent sets of ground literals, closed with respect to inference using RULES.

Let TREE be the set of all the nodes in the proof tree that become true when the action in question is performed.

Let CHANGES be the set of facts entailed by TREE U NOW but not by NOW alone. ("|-" and "|-/-" are "entails" and its negation).

CHANGES = { S | NOW U TREE |- S & NOW |-/- S}

If this set is inconsistent, the updating process is undefined.

Let SAME be those facts true in NOW and left the same by the action, and NEXT be the SOA that follows the update.

SAME = { S | S member NOW & -(-S member CHANGES) }

NEXT = CHANGES U SAME

NEXT must be consistent if CHANGES is. Proof:

If x member NEXT, then x member CHANGES or x member SAME

If x member CHANGES, then NOW U TREE |- x

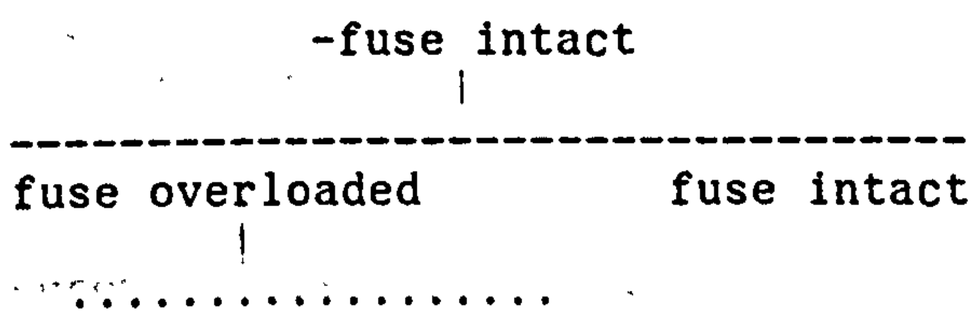
If x member SAME, then x member NOW, so NOW |- x, so NOW U TREE |- x

So if x member NEXT, then NOW U TREE |- x

So if NEXT is inconsistent, CHANGES must be too.

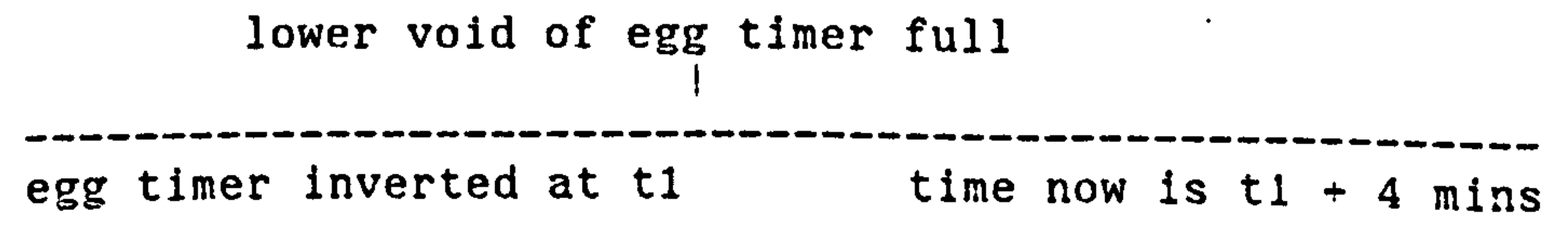
4.3.1.1. Events are like actions

This gives me a mechanism to define "event" that points up the likeness of events and actions. An event, such as a fuse blowing, is the becoming true of a proof tree of some particular sort. For instance, the event "fuse blows" would be the coming true of a proof tree like



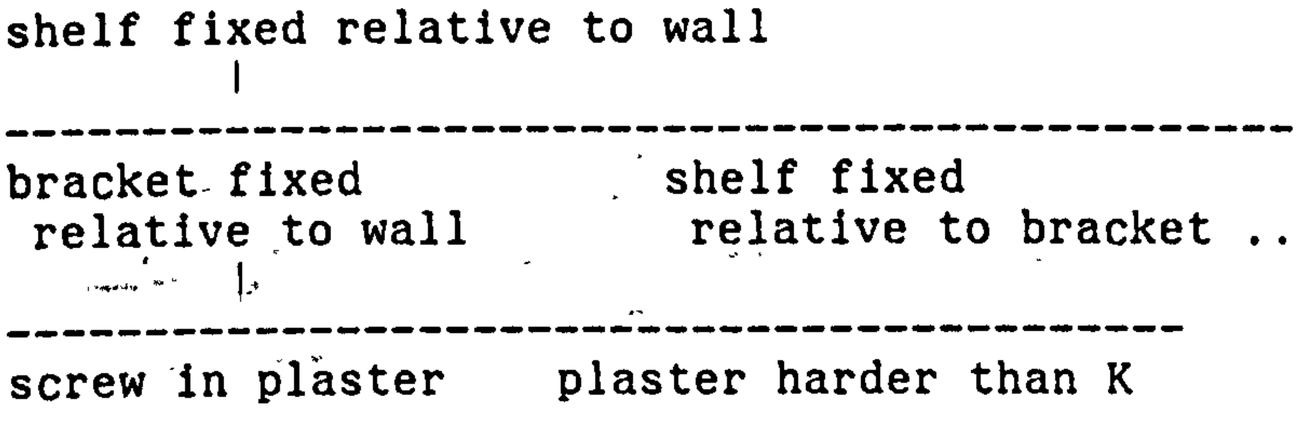
The way RULES + NOW + EVENT-TREE (the facts in the proof tree that defines the event) update NOW will be just the same as for actions. An action takes place when the simple actions at its fringe become true, because of what the agent chooses. An event is the same, except that the facts at its fringe that become true are not actions.

Why do these fringe facts become true? I could just say that at some level the universe "just does" those things, but before I could say that I would have to have broken the proof trees down to a level where their fringe facts were at the limits of physics. At any higher level, there is an answer to "Why?". The right answer (not vital to me) may be like this. A SOA must be described, not just by the facts in it, but also by an attached date. Some rules true in the SOA will be sensitive to this date. Then the event, "egg timer runs through" would be the coming true of a proof tree like this.

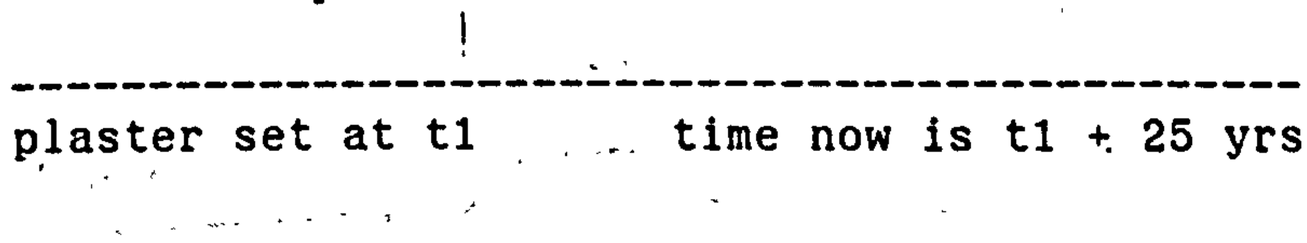


(Clearly one might use more rules to give a more detailed account of the event. But the principle of facts true because of the date at which the proof tree is considered will stay the same).

There are also events that appear to need to be described by saying that a proof has ceased, rather than begun, to be sound. One could define "the shelf on the wall collapses", occurring as the wall softens, as the ceasing to be true of a proof tree like this.



Then the decay through time of the plaster could be caught by a rule such as



When eventually this rule fires, the update process will lead to the deletion of "plaster harder than K" and the collapse of both the proof and the shelf.

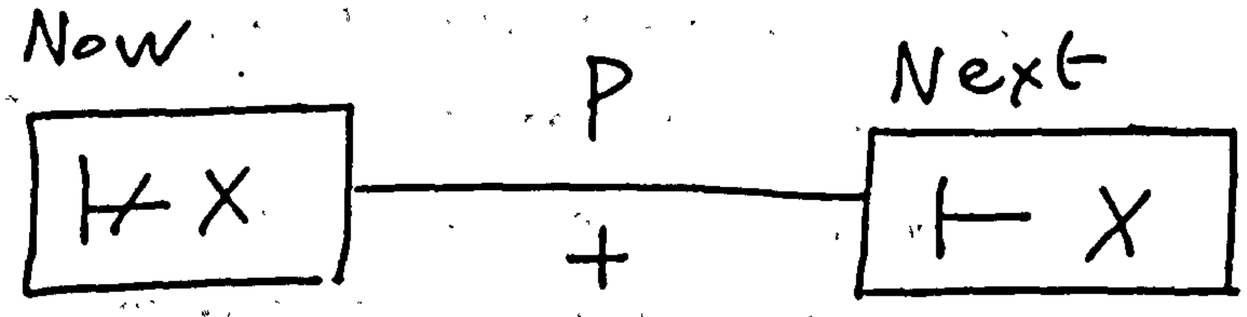
4.3.2. Expectations and choice

4.3.2.1. Expectations

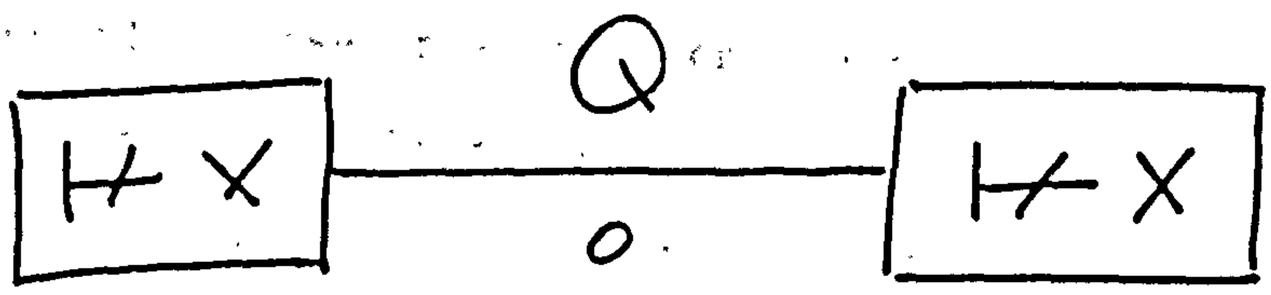
Expectations are drawn as in Expec/1. A SOA is drawn as a rectangle. Its name is written above it. What is or isn't provable in it is written inside the box.

Time runs from left to right. The arc between SOAs can be labelled

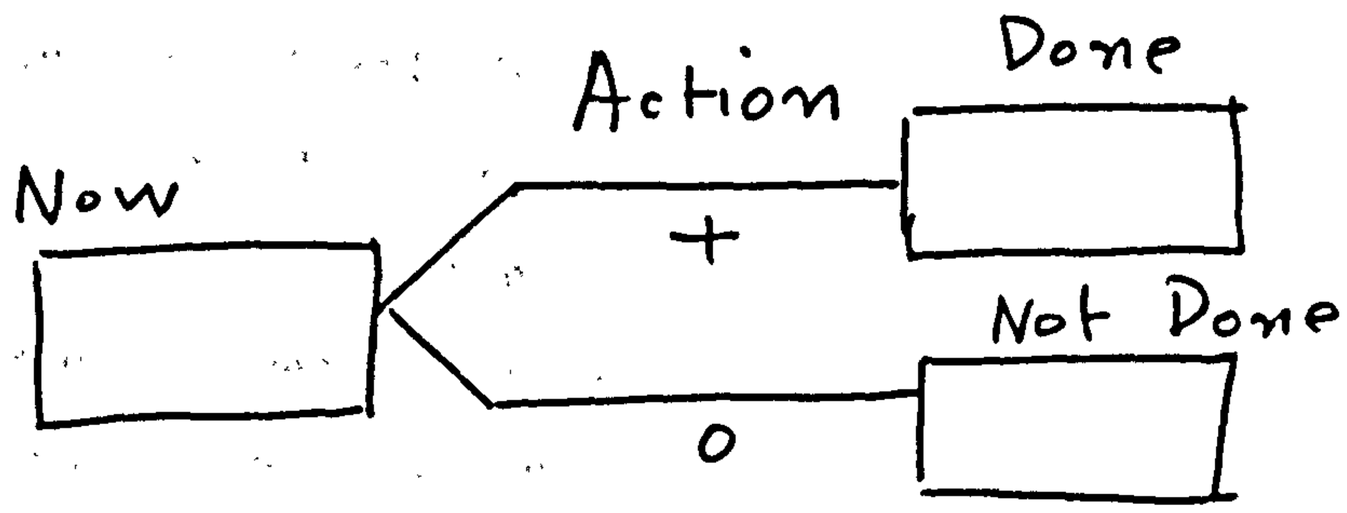
Expec/1



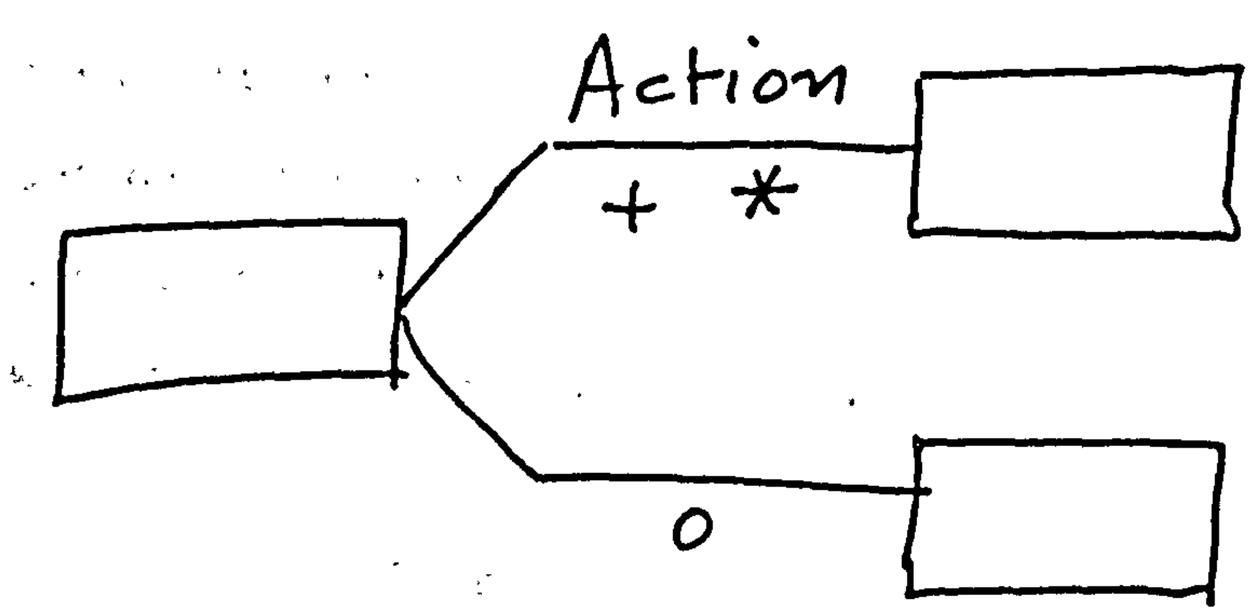
Expec/2



Expec/3



Expec/4



with (the name of) a proof tree. If the preconditions of the proof tree P are true in NOW, the arc labelled P can be labelled with a + too, to recall that the nodes of P have to be added to the first SOA to get the next one. If the preconditions are not all true, the arc can be labelled with a 0, as in Expec/2, to recall that NOW should be copied without alteration to derive NEXT.

Whether the preconditions of the proof tree of an action are true in a SOA is not the only determinant of whether that proof tree is sound. Those preconditions that are simple actions are true at the election of their agent. So perhaps the nodes of the tree have to be added to get the next SOA, perhaps not. This is represented as in Expec/3. When the agent chooses whether to act or not, he then knows what will follow. His choice is marked by a star on one arc, as in Expec/4, where he has chosen to act.

In Expec/3/and Expec/4, and some later diagrams, what is being illustrated is not any particular possible course of events, but classes of courses of events which have in common only the shape of the choices open to an agents and the choices that he actually makes among them. To emphasize this the boxes representing SOAs are left empty of any assertion about what will or won't be true in that SOA.

4.3.2.2. Choosing what to do

When you contemplate the execution of a plan, you are faced with an option. You can turn the world as it is now into either the Done world, as it will be if you act, or into the NotDone world, as it will be if you don't. Given this option, you make a choice. How?

Assume that there is a function Judge which takes a world, finds all the true good facts in it and counts their worth, finds all the true

bad facts in it and counts their worth and subtracts the good from the bad. Given a plan, the simple benefit of doing it could be described as

Judge(Done)

The simple benefit of not doing it could be described as

Judge(NotDone)

But contemplation of an act is not all that is involved. You have to perform it, and that takes effort. Some great goods take so much effort that they will not be done. Imagine another function Effort which measures the effort involves in executing a plan. The merit of performing a plan will be

Judge(Done) - Effort(plan)

The merit of abstaining from it will be

Judge(NotDone)

(In fact you may have several options, each with its own merits. But the option of doing nothing is always there.) These are the apparent merits of your set of options. You want the best future SOA. So you choose the action (or abstention) that seems to lead to it.

Your choice is not truly between the possible SOAs may that result from your possible actions. You have estimated the outcome, but you know you may be wrong. What you really choose between are your possible actions. This distinction is important.

4.3.3. The effects of what a speaker says

Your choice can only be made on the basis of what you know about the Now world, and what you value. That is what tells you whether your plan will run to completion and what consequences it will have. And you can assume that your current values will remain those that you judge the future worlds by.

Unless of course they are changed. One could distinguish the changes that arise from what you perceive of the world from those that arise from what you are told about it. But when one is interested in teleological explanation, a better division is between those updates that arise from your actions (such as looking to see, or listening to hear), and those that arise from the other's actions. These are typically verbal, but need not be - eg the display of a cut finger; or of a broken rope to clinch an argument about whether a particular method of suspension was wise. There will be many assertions true of such a non-verbal "remark". Explaining it as if it were an utterance will require taking one of them and showing that if Hr comes to hold that assertion, then Sp benefits.

I am distinguishing between facts and values. Both can be expressed, but they are quite different and affect Hr by separate mechanisms.

4.3.3.1. The effects of expressed fact

I talked at some length about how remarks could have any effect on the hearer at all; in particular, about how they could ever be taken to be an attempt to present a picture of the world. I assume here that this can be done.

But presenting a picture of the world is not the same as presenting a

picture and claiming that the picture is veridical. I suggested that an expression of fact first had to be seen as sincere.

4.3.3.2. Opinions about fact

When I talk about an attempt to changes someone's opinion about a matter of fact, I could mean any of 3 things.

An attempt to change their belief about its truth

Or about its provability

Or about a particular proof of it.

I suggest that if one wants to talk about language use, the last is the most important.

Clearly the first is not useful. Humans do not have access to the truth. (Indeed, I would argue that truth is merely a reification we use to explain the oddity that at t1 we hold that at t1 F is true, but at a later moment t2 we hold that at t1 F was false.)

Sometimes though we believe that we know enough about the world to bet that we won't be surprised this way. Then we can make the closed world assumption; to wit that |-/- X entails |- -X. But in the real world we often know we are ignorant, and daren't make the bet.

Logically, the second must be what I mean. At least, if at the end of an argument you no longer hold that P is provable, I have convinced you. And if you still hold P, I have not. But during the argument, I have to describe what I do in terms of attacks on specific proofs. Suppose you hold that nuclear disarmament is a bad idea on the grounds of Soviet agressiveness and of reduced research funding. Suppose I attack you on the first of these, which I believe I can

convince you over. But suppose at the same time I think that I don't have the facts or persuasiveness to convince you of the second. I know I will leave you thinking that nuclear armaments are worth having. Does this mean my action is irrational, since I know it can't succeed? No, if you explain me as attacking a particular proof. You can say that though I don't see how to finish your conversion, my attack on one proof tree of your belief at least tends to the goal I seek.

Similarly, suppose you are refusing to accompany me to the cinema, because you will be bored both by the tedium of the drive there, and by the tedium of the film. I tell you how well reviewed the film has been. I hope that if I convince you, the tedium of the drive will not stop you going. But I still know that you will still find it provable that you will be bored. So my action is explicable only as a an attack on a single proof of a proposition, not on its general provability.

4.3.3.3. The effects of expressed value

The obvious way to express a value is to say "I want X" or "I don't want (=fear) X". If I say this, you can doubt my sincerity, but not my authority. I am the expert on what I want.

Insofar as my wants are objects of your belief, my expression of them is expression of fact just like any other. But it can be more than that. It may be that you will let you own values be influenced by mine. General benevolence is the obvious case. If you are benevolent towards me, then just your knowing that I want something will instill in you the desire that I get it, for no other reason than that I want it. General malevolence is also possible: if you know I want something, thereupon you want me not to have it.

I can also affect your values by expression of other people's values. I could say "Celia wants to be met at the station". Here I am not authoritative in the way that I am about my own values, but nevertheless you may accept my authority on other grounds - for instance, that I have just spoken to her on the phone. If you believe me, and if you are benevolent to her, you will adopt her reported value as your own and go and meet her.

Adoption of other people's values does not have to be general. It can be true for just some sorts of goal. For instance, parents who practice demand feeding adopt their child's goals if they are that it be fed, but not if they are that it play with a power point. It is not the case that that they do it in order that the child will not starve. Feeding it at any time would do that. Feeding it when it wants is feeding it because it wants.

4.3.3.4. Accepting belief and accepting value are alike

There is a likeness between the way a hearer may infer what a speaker believes, and what a speaker values. In the case of belief

> Hr observes that: Sp says Sp believes X
>
> so if Hr believes Sp is sincere, Hr infers: Sp believes X
>
> and if Hr believes Sp is authoritative, then: Hr believes X

Similarly, in the case of value

> Hr observes that: Sp says Sp wants X
>
> so if Hr believes Sp is sincere, Hr infers: Sp wants X
>
> and if Hr is cooperative towards Sp, then: Hr wants X

4.3.4. How Sp benefits

Suppose Sp's utterance does change Hr's choice about his intended action. Why is this a good thing for Sp? Sp's action can only be explained by benefits he expects to receive.

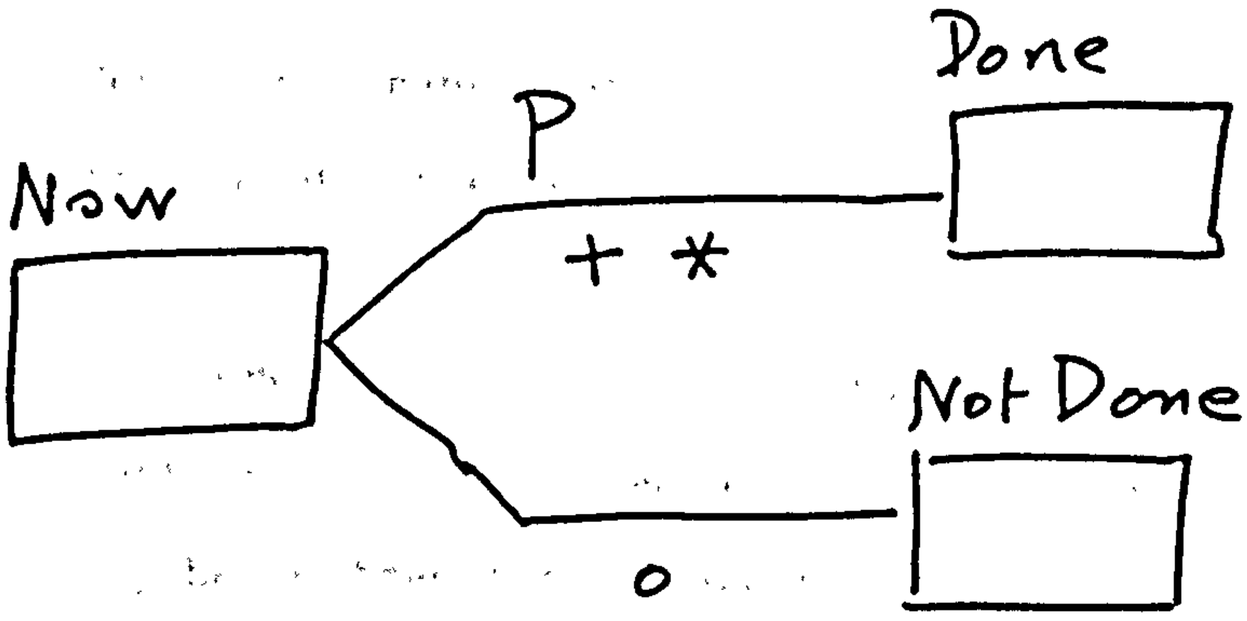
His benefit is not that he has changed Hr's choice. Mere deceit might do that. Nor that he has changed it, as he thinks, for the better, so that Hr is working with a better picture of the world and is therefore more likely to succeed in what he attempts. If that were so, it would be impossible to explain why Sp should ever attempt to deceive Hr, so that Hr acted in a way that Sp thought more likely to fail.

No. What matters is that when Hr makes a choice, he choose a course of action. Later he will execute it in the real world, which he and Sp share. What Sp expects as the result of what Hr does depends on how Sp thinks the world really is. How much Sp likes what he expects depends on what he really wants. What Sp expects to follow Hr's action need not be what Hr expects.

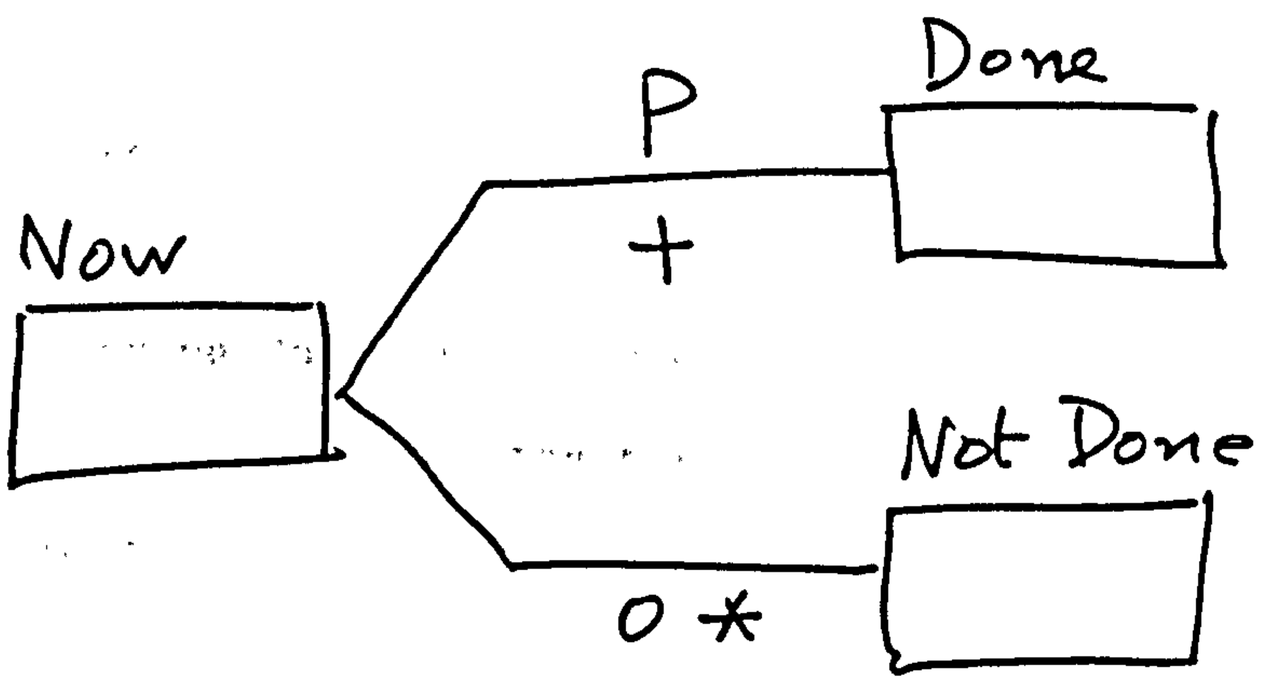
To see Sp's benefit, look at the world after the choice Hr would have made if Sp hadn't spoken, then at the one that follows the choice Hr will now make, and see why the latter is better for Sp.

Initially Hr's choice might be as in Alts/1. Then Sp speaks. Hr's beliefs and values are changed. He re-evaluates what he is going to do. His choice is now as in Alts/2. But now look at that from Sp's point of view. Sp's action is his utterance. The worlds that result from his utterance differ only in what Hr thinks, but different they are. Sp's options are as in Alts/3. Sp's choice is governed not by the immediate result of his utterance, but by what flows from that

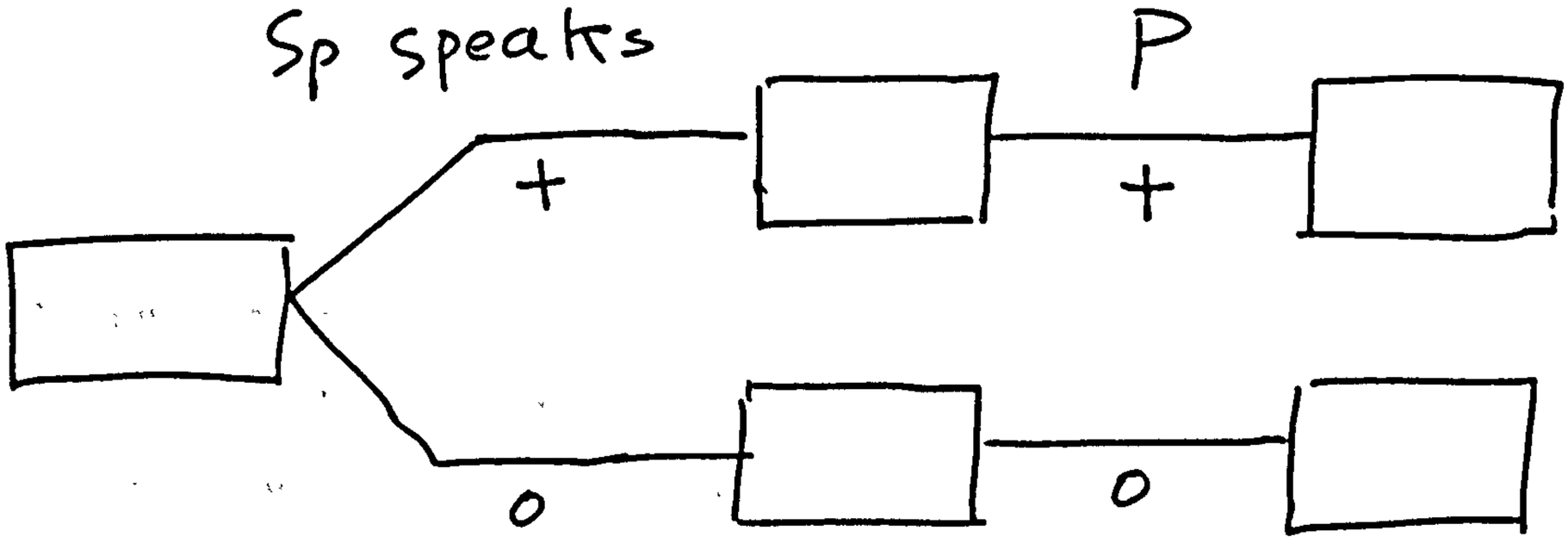
Alts/1



Alts/2



Alts/3



after Hr in his turn acts.

Could one represent this with an ordinary game-tree, such as Alts/4? No. The differences are that

- There may be no description of the world at the nodes of the tree that Hr and Sp agree about. If they expect different results, there may be no tree that shows both their actions as rational.

- In a game explained by a game tree, one player moves and waits to see what the other will do. But after Sp has spoken, he does not see himself as waiting to find out what Hr will do. He believes he has forced Hr's choice. Hr's reasons for acting are part of a different tree.

Finding Sp's benefit has two twiddles to it. One is deciding in which SOAs to look for the benefit. The other is that Sp need not be purely selfish.

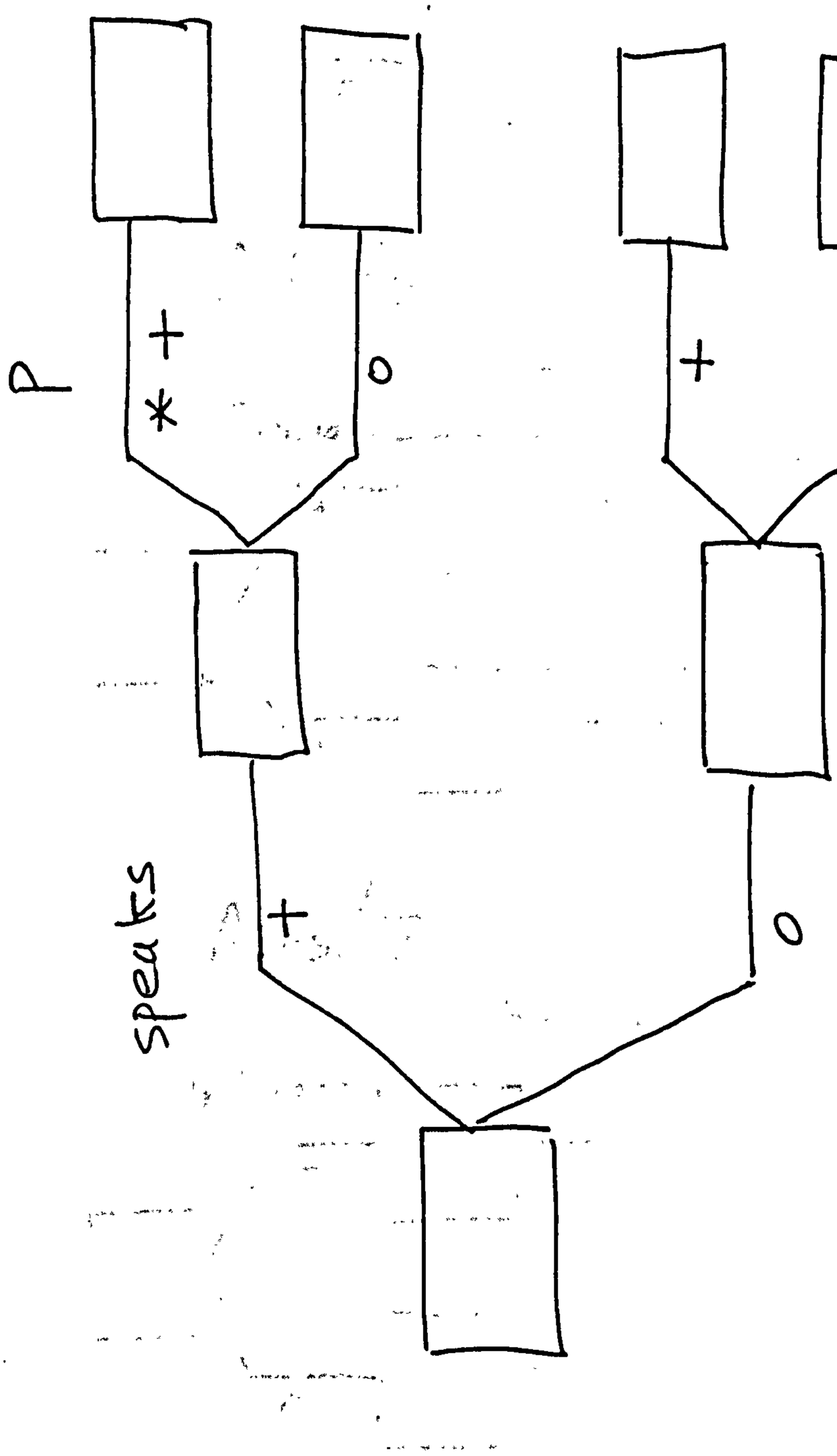
4.3.4.1. Three moments to benefit

Once Sp has spoken, we can look forward various distances into the future to see how he benefits.

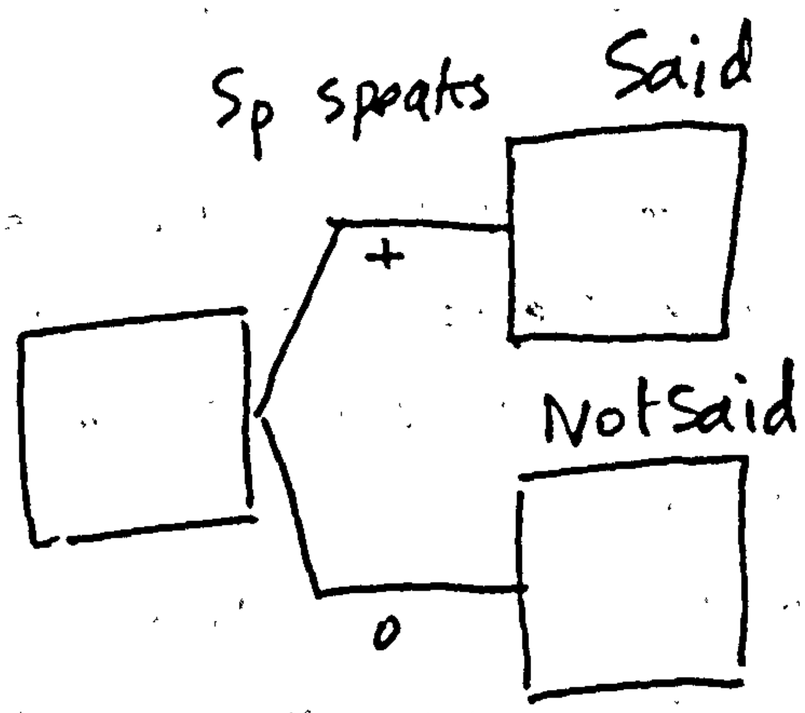
4.3.4.1.1. Hr's beliefs and values may change

In abstract argument about such things as the good, the beautiful and the true, I may seek to persuade you, not so that you will act differently, but just in order that you believe differently. To see the benefit to me in this, we need look no further into the future than is shown in Alts/5. Any benefit to me will accrue merely because I would rather that you had the beliefs that you have in Said than

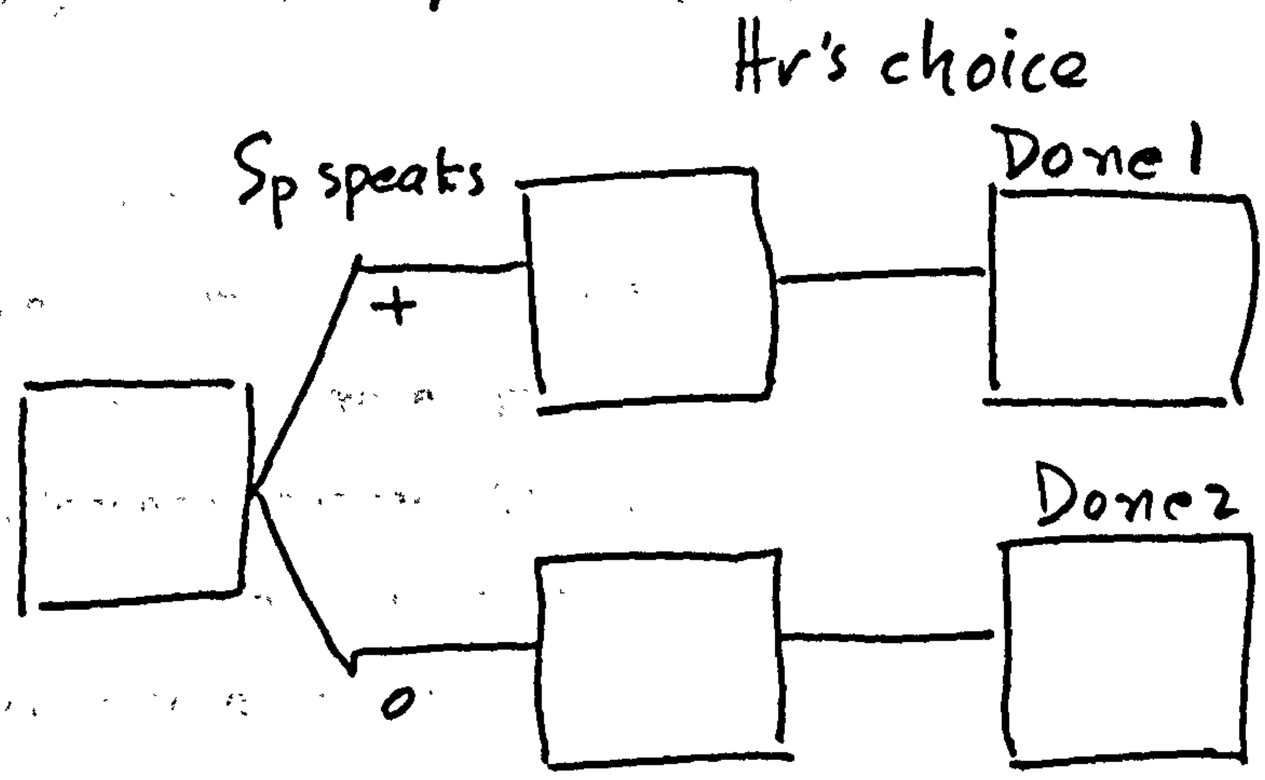
AIIs/4



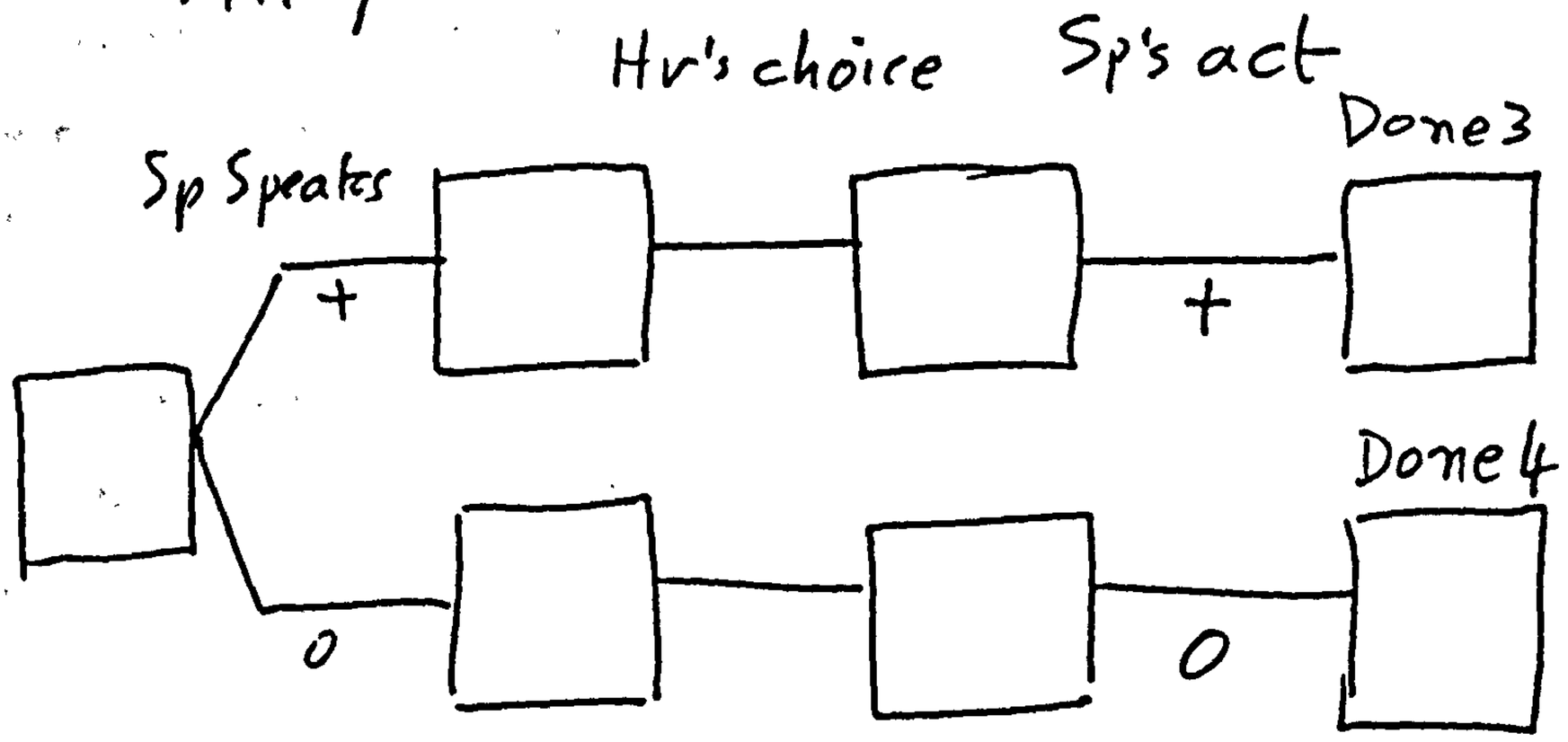
Alts/5



Alts/6



Alts/7



those that you have in NotSaid.

4.3.4.1.2. Hr's actions may change

As a result of what you come to hold, your chosen action may change. If it does, the worlds I expect after you have executed your choice will be different. One needs to look as far ahead as is shown in Alts/6. The difference between Done1 and Done2 may be my final goal. It will be if for instance I have advised you that the shops are shut so that you give up your abortive plan to go shopping; or if I have asked you to turn down the heating because I am too hot, and you do.

4.3.4.1.3. Sp's opportunities may change

And lastly, if you act differently, my opportunities for action may change. Once you have acted, you leave the world in a state where my actions are made possible or easier. One must look ahead to the difference between Done3 and Done4 in Alts/7 to see my benefit. This arises when for instance I ask you to lend me your bicycle. You choose to do so; and I can then make my journey more easily.

4.3.4.2. Other peoples values affect Sp's

What counts as Sp's benefit need not be a matter of pure self-interest. There are three basic ways that Hr's values can be related to Sp's.

- Sp can be benevolent to Hr. If Hr wants X, Sp wants X.
- Sp can be malevolent to Hr. If Hr wants X, Sp wants -X.
- Sp can be indifferent to Hr. Hr's wants do not affect Sp's.

The criterion that makes Sp's utterance rational is that it leads to

a better SOA for Sp. But what exactly this is may depend on Hr. And similarly, Sp may cost Hr's effort as his own. When Sp advises Hr, he is doing exactly that.

4.4. A simpler way of choosing what to do.

I've presented a picture of how one chooses what to do which involves computing a picture of the whole world as it will be after the action and contrasting it with another picture of the entire world as it will be if the action doesn't take place. This may be theoretically elegant, but as a practical way of making choices it must be a dead loss. Any process that involves transitive closures under inference will be.

A more reasonable process would start with the action, and look out to see what effects it does or doesn't have, regardless of how the rest of the world is. An action should be chosen if and only if its effects are on the whole beneficial. Changing a person's choice about his a plan is a matter of changing his opinion about what its effects are and what their merit is.

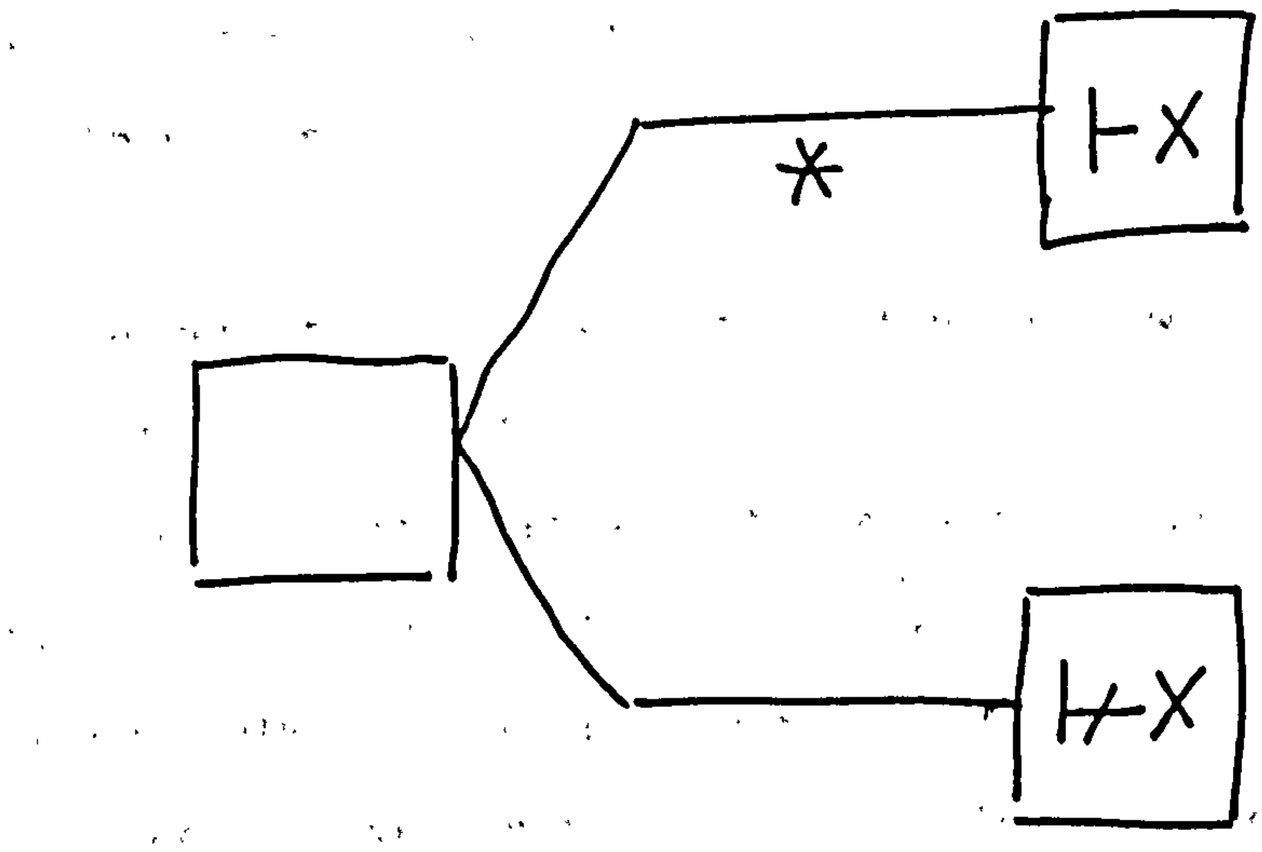
But what is an effect? And what is an effect an effect of? Consider Sns/1. In i, X is provable after the chosen, starred, course of action is executed. Otherwise it isn't provable. One could call X the effect of the choice, since it occurs if one does what one has chosen to, but not otherwise.

In ii, X occurs after some action P is done, but not before. Perhaps X is the effect of the action P. Though it need not be. X may happen anyway, as in iii.

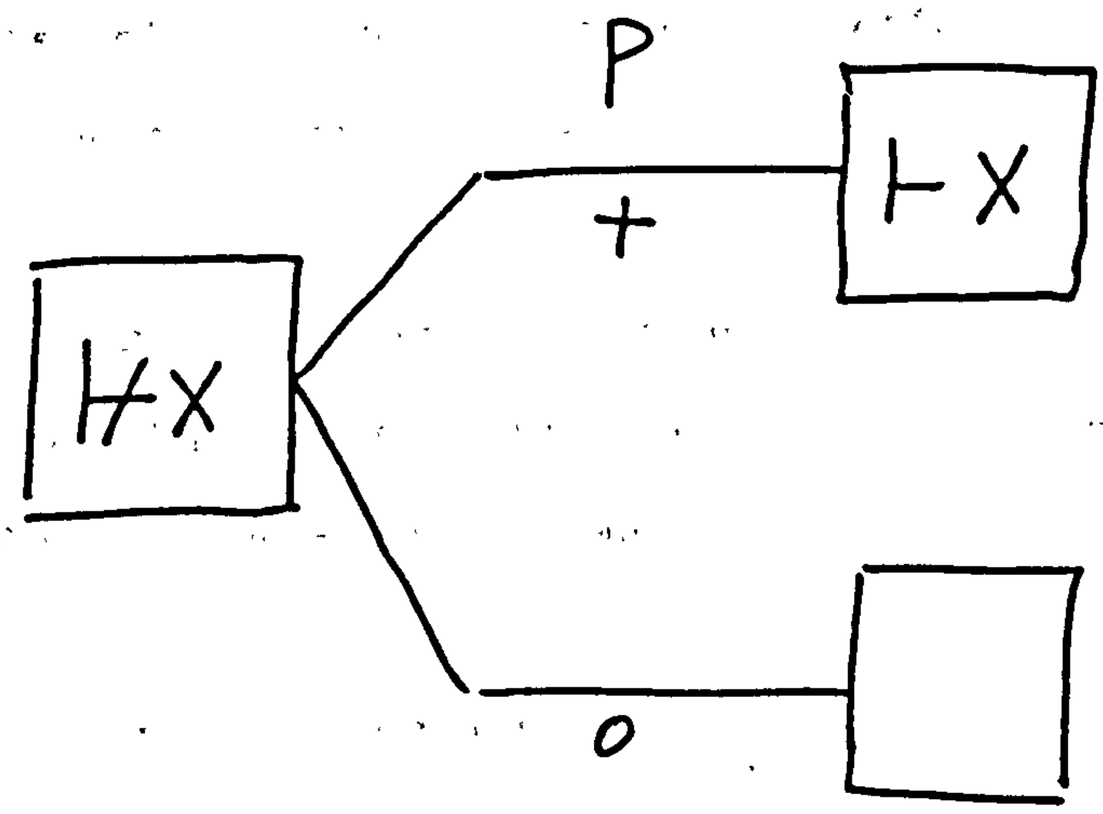
The effect that matters is the effect of choice. I say this because

5ns/1

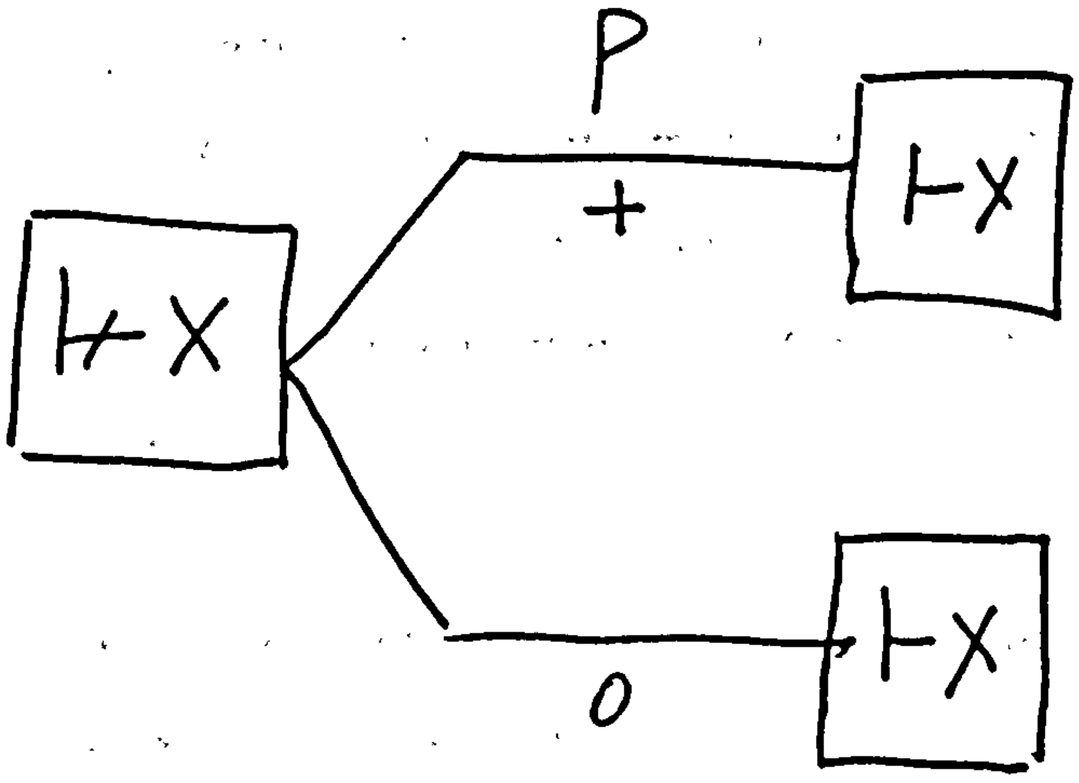
i



ii



iii



if one's expectation are as in iii, where there are effects of action but not of choice, one's anticipation of X is not going to affect one's choice. How could it? It is the same whatever one's choice. Something else, or an arbitrary choice, must sway one.

But given that one is contemplating a known plan, it is much easier to find out about the effect of that action. And indeed this will often be as good as knowing the effect of choice. The reason is that, typically, if nothing is done the world stays the same. Sns/2 shows a choice between doing P or doing nothing. If one does nothing, and if one can assume that nothing is going to happen spontaneously, then the worlds Now and NotDone are the same. So contrast of Now and Done is equivalent to contrast of Done and NotDone. By looking for the effects of action one finds the effect of choice. What follows is about how this can be done.

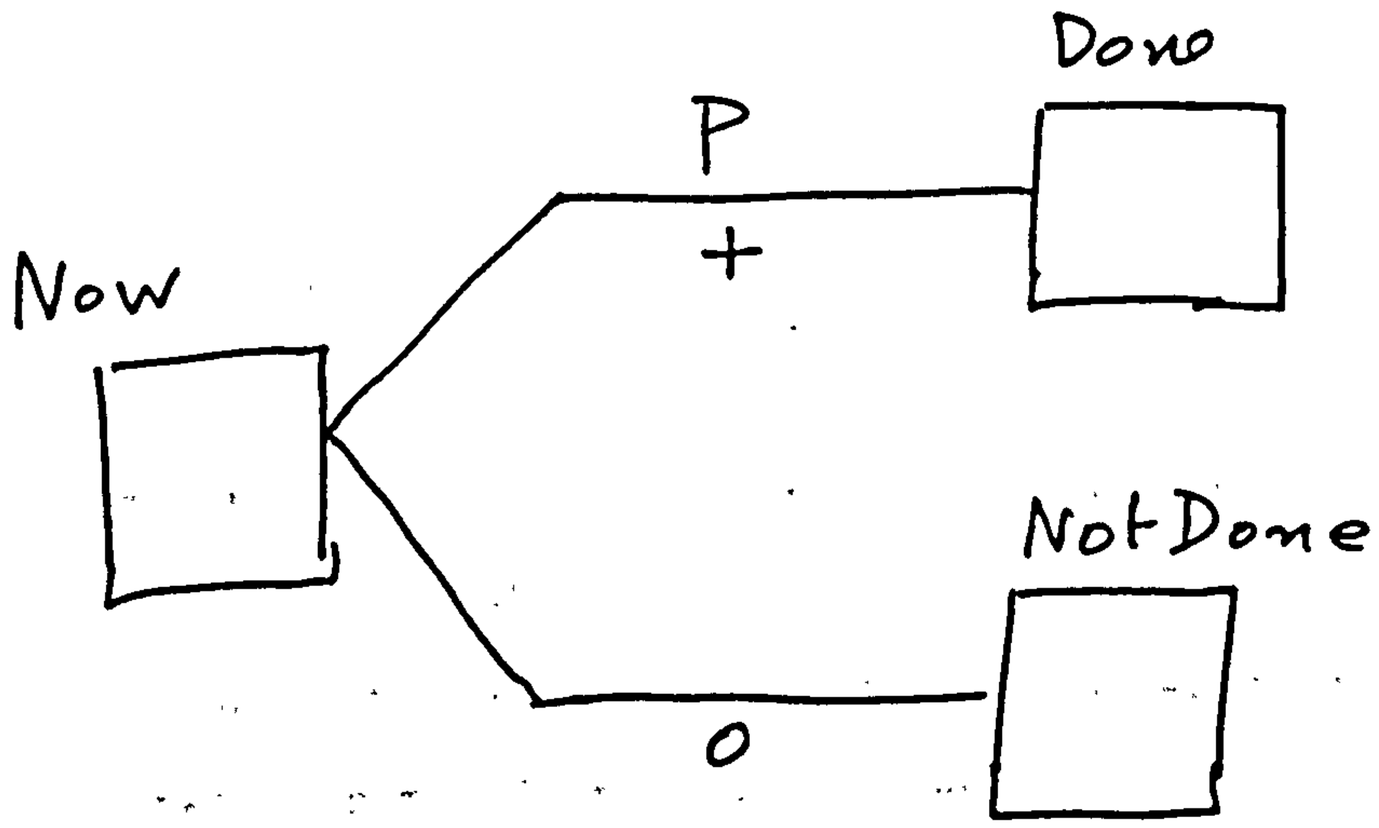
So what are the effects of an action? Or more precisely, what are the effects that matter to your choice? For V to be an effect that matters, it seems that it must be

- provable after you act
- not provable before you act
- valued; either good or bad, not indifferent

If these are true, and the assumption of no change holds, then you have the expectations shown in Sns/3. If V is better (or worse) than its absence, then Done should be preferred (or not preferred) to NotDone. One has derived the effects of a choice from the effects of an action.

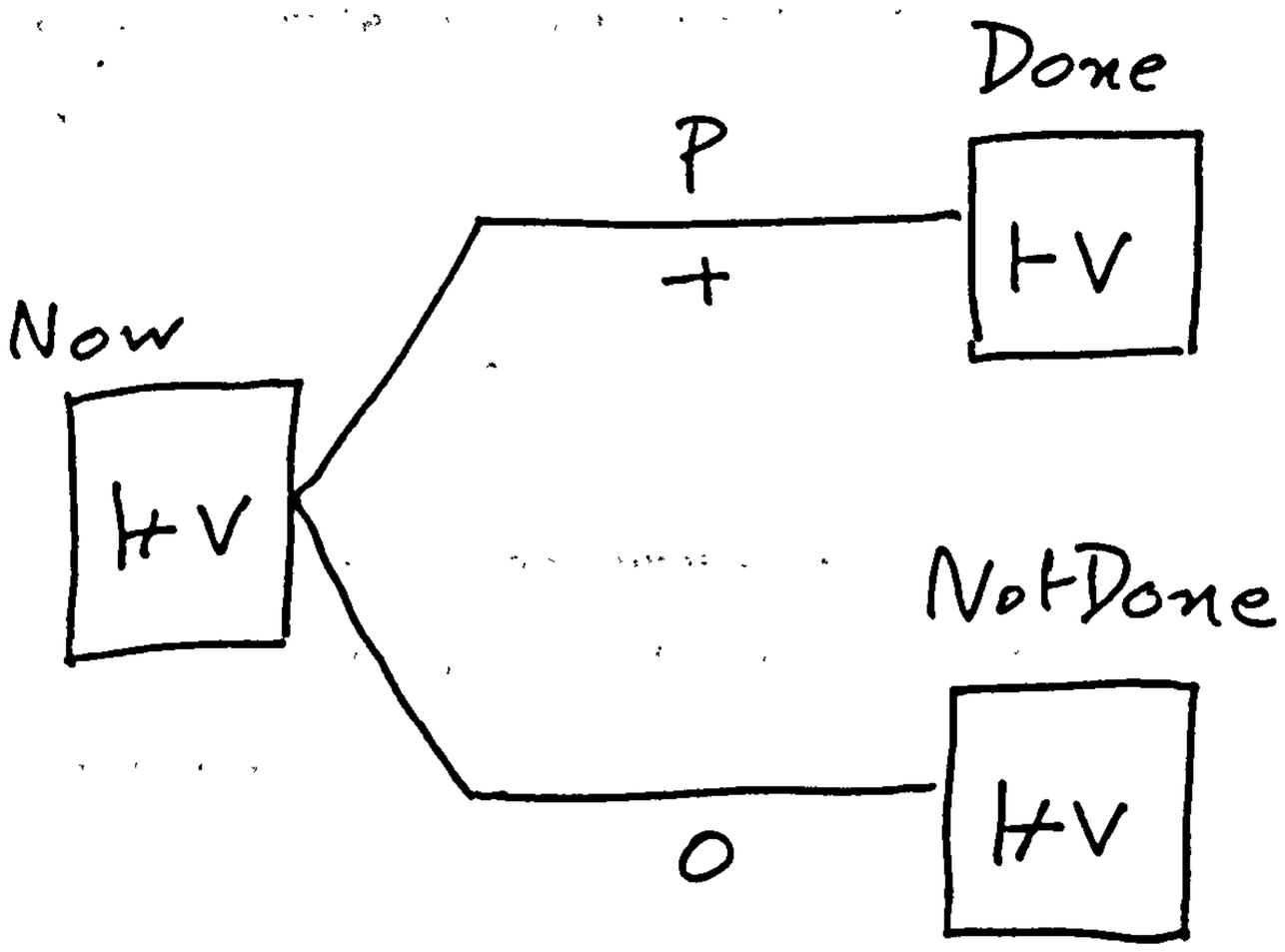
Why are the effects of action easier to find than the effects of choice? Suppose the contemplated action is

Sns/2

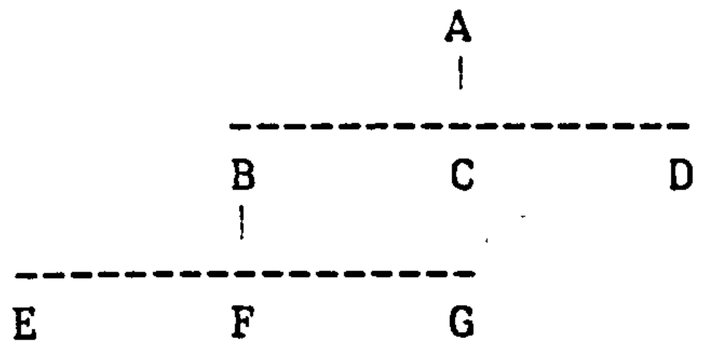


Now = NotDone

Sns/3



TREE 1

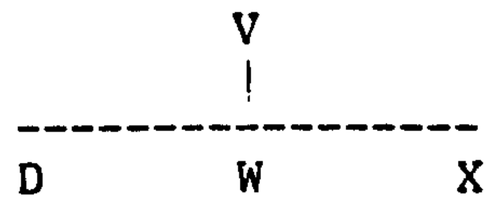


To find its effects one can search forward from those facts that must be true if that action is performed, (the nodes of the tree) and see if one can construct a proof to a valued fact (V, say). By searching forward from a fact, J say, I mean

looking for a rule C <- J & L1 & ... & Ln,
if so, seeing whether one can prove L1,...,Ln,
and if so, deducing, and searching forward from, C.

What one is doing is finding a member of the set of facts CHANGES, so those facts provable in NOW, the initial SOA, are also available to fire the rule. Perhaps one finds a proof of V like this:

TREE 2

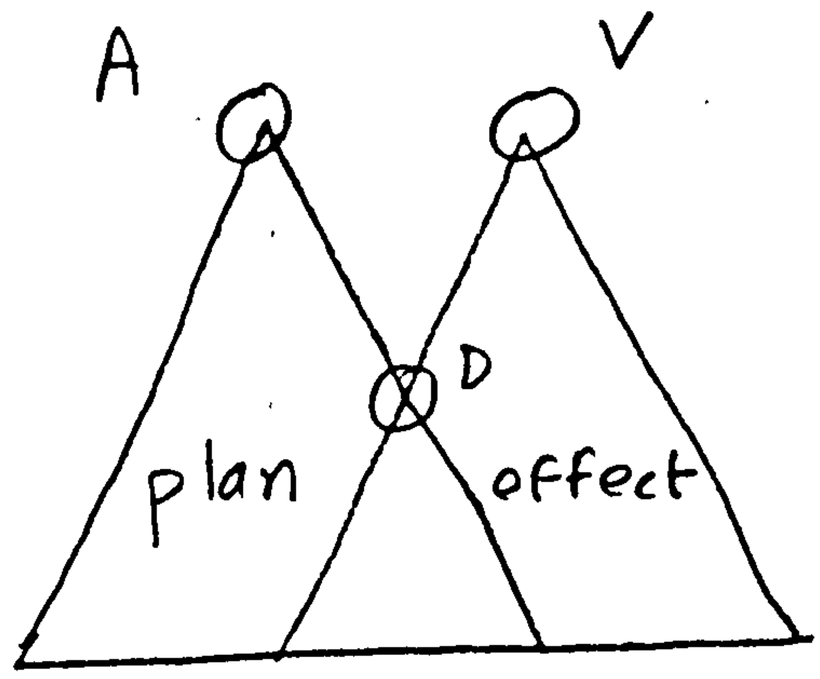


where W and X are provable in NOW. If it then turns out that it is impossible to prove V in the SOA before one acts, then V is an effect of the act.

This can be drawn as Sns/4, using the conventions described in the appendix. The overlap represents the occurrence of D in both of the plan and the effect proof trees. The triangle below D is the sub-proof that the proof trees share.

Searching forward from the nodes of the contemplated action does not

sns/4



guarantee that the valued facts one can find are effects. They may also be provable in NOW alone. But if one does not find them by searching forward from the action, then they definitely are not effects. Either they are not provable after the action, or they are, but were provable in NOW as well.

Such search could be continued indefinitely, but must in practice be limited. But this is a separate matter to deciding what the search is for.

I make a distinction between the necessary and the accidental (or side) effects of an action. Necessary effects are those that are entailed just by the facts in the proof tree of the action. For instance, when I buy something, this has the necessary effect of my having less money afterwards. This happens whatever the initial state is in which I do the buying. But some effects only arise because of the circumstances in which I do them. I will get wet when I go out only if it is raining, not whenever I go out.

The goal for which a plan was made should be a necessary effect of the plan. A plan to get eg "have milk" should be co-extensive with the proof of its good effect "have milk".

4.4.1. Undermining proof trees

Actions can add valued facts to a future SOA. They can also take them away. Suppose the water in the kettle is boiling because it is on the hob. This proof tree is sound.

water in kettle is boiling
|
---
water in kettle kettle on hob hob above 100C

Then the action of taking the kettle off the hob, which might have a

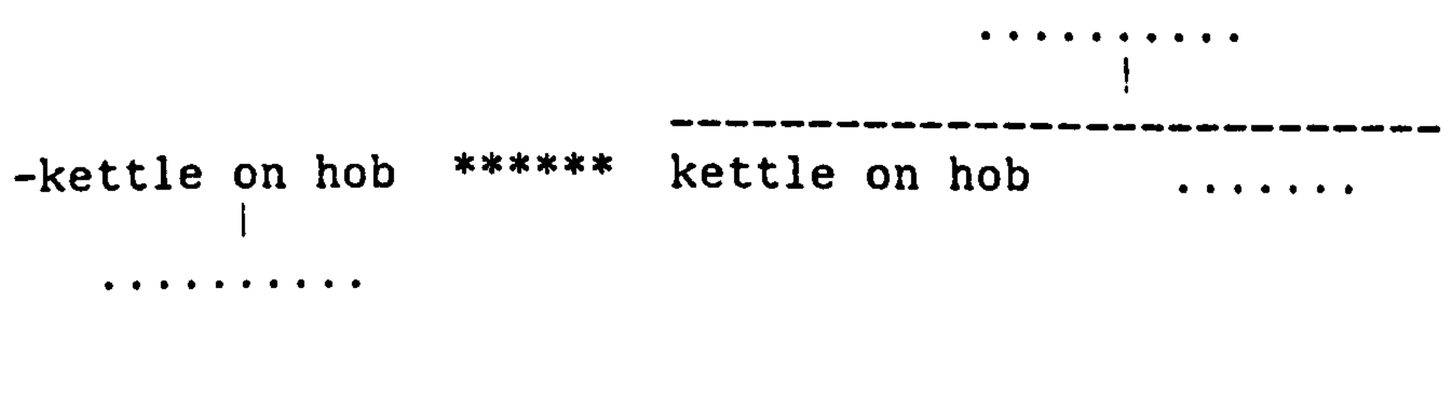
proof tree like this

- kettle on hob

... various hand movements ...

is going to disrupt the first proof tree. Before my action "water in kettle is boling" was provable; after my action, it is not. That should count as an effect of my action. But according to the criteria above, it doesn't. Those criteria require something to go from being unprovable to being provable. Nevertheless, sentences ceasing to be provable also matter for making a choice. The SOA diagram of what will happen is Under/1. If the no-change assumption holds then the change that the action brings about is an effect of the choice, as in Under/2.

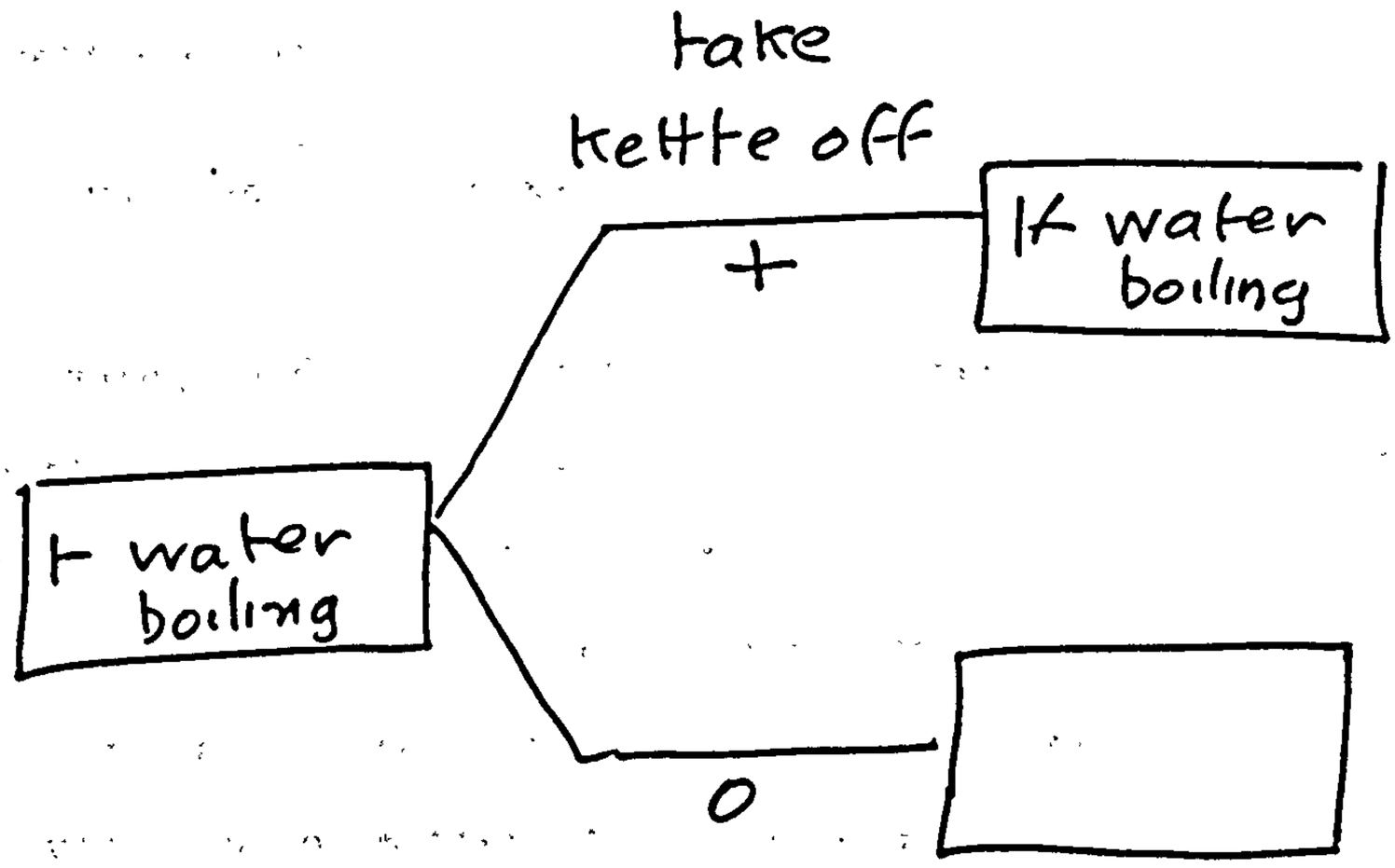
In the shorter process, I must eschew comparison of SOAs, and instead look at how proof trees interact. Undermining can be handled in this way too. Rather than the action supporting another proof, one of the action's nodes (here the topmost) contradicts one of the preconditions of the proof tree that is being undermined. I draw this relation as Under/3. The square marks the contradiction, such as



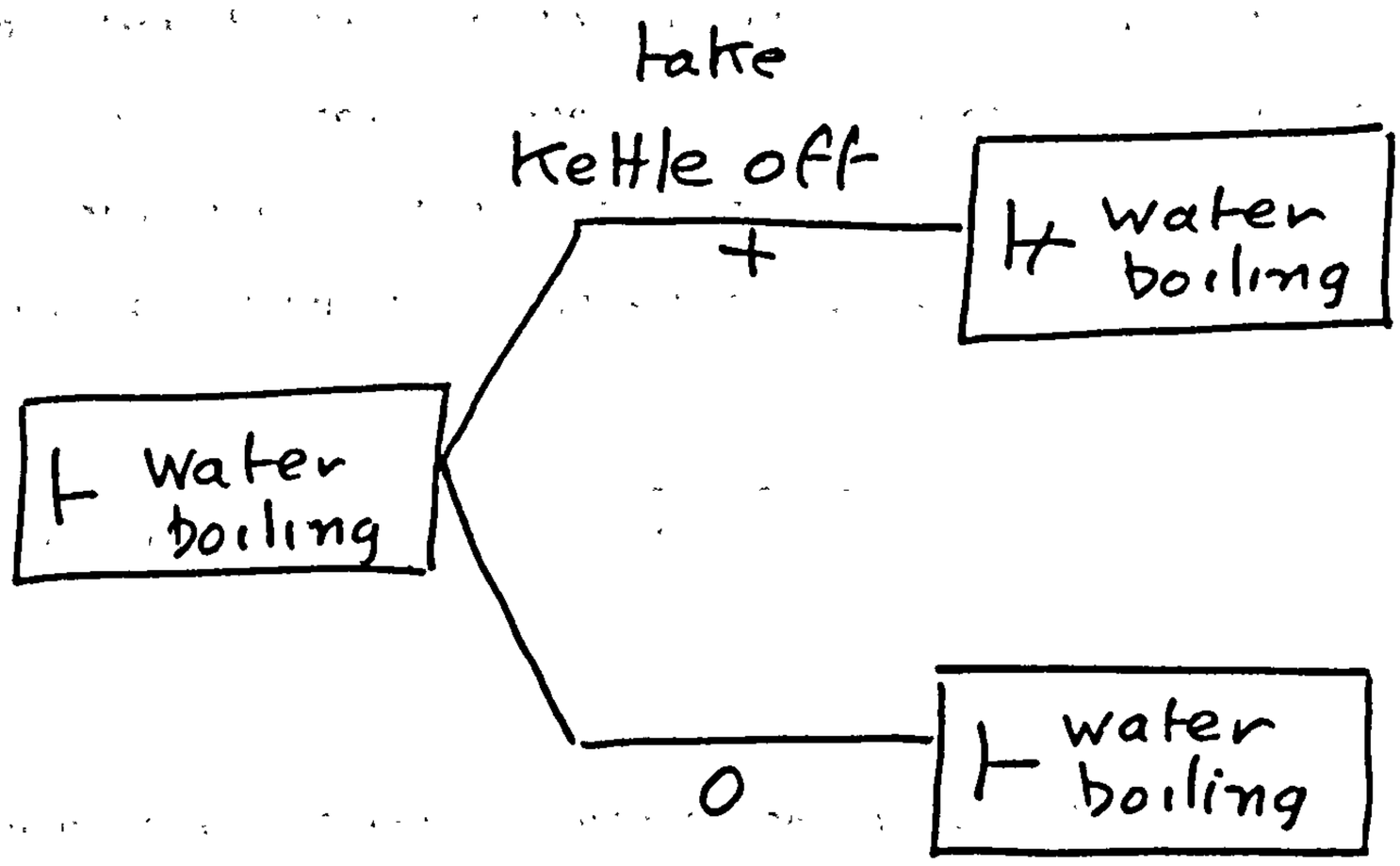
No overlap between the proofs can occur. They are contradictory.

If an action makes a proof sound, the merit of the plan is increased or decreased by the merit of the consequence of the proof. But if it undermines a proof, the merit is changed by the opposite of the merit of the consequence. Undermining bad is good and preventing good is bad. Showing this on SOA diagrams is easy. Look at Prev/5. If E is a

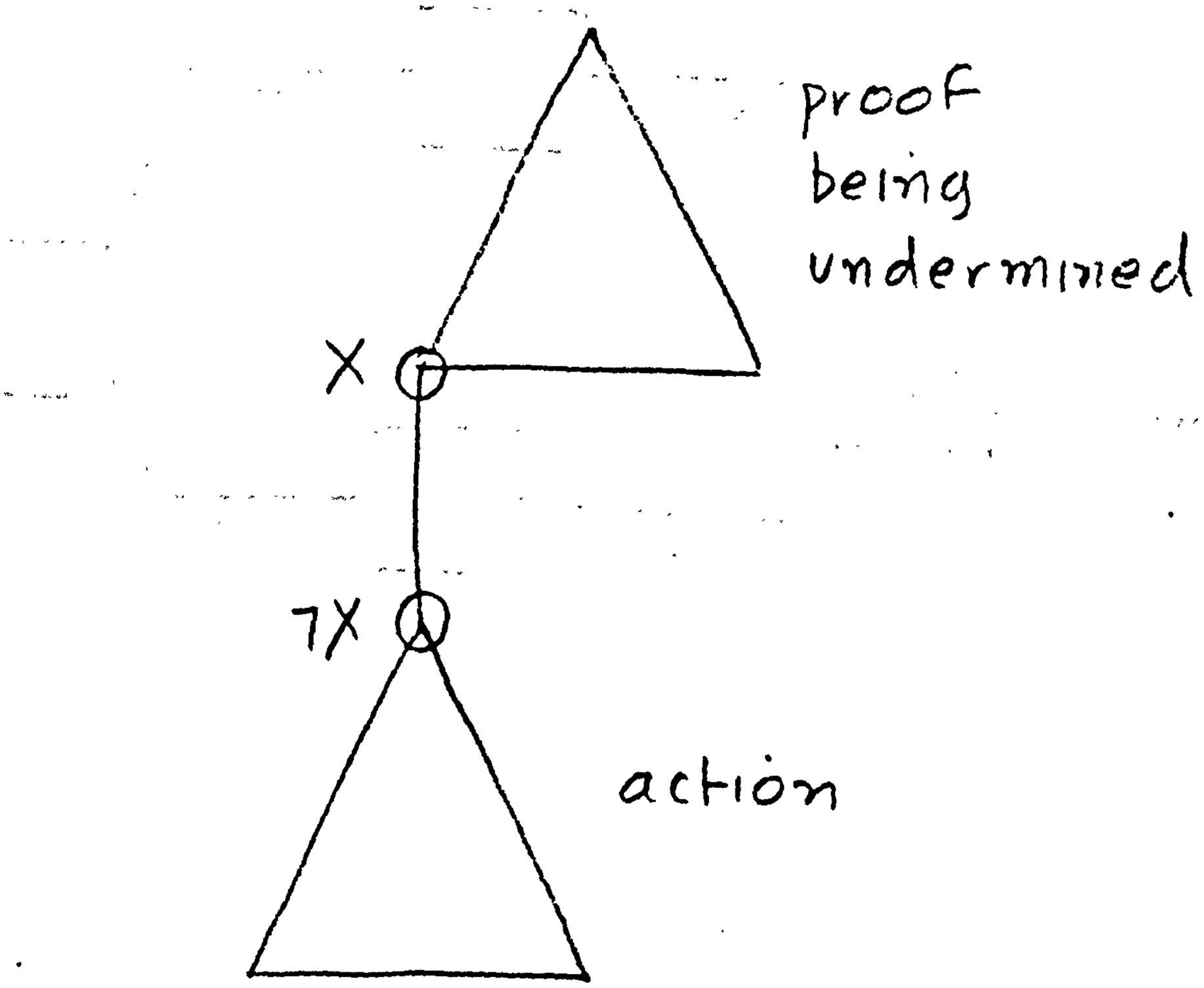
Under/1



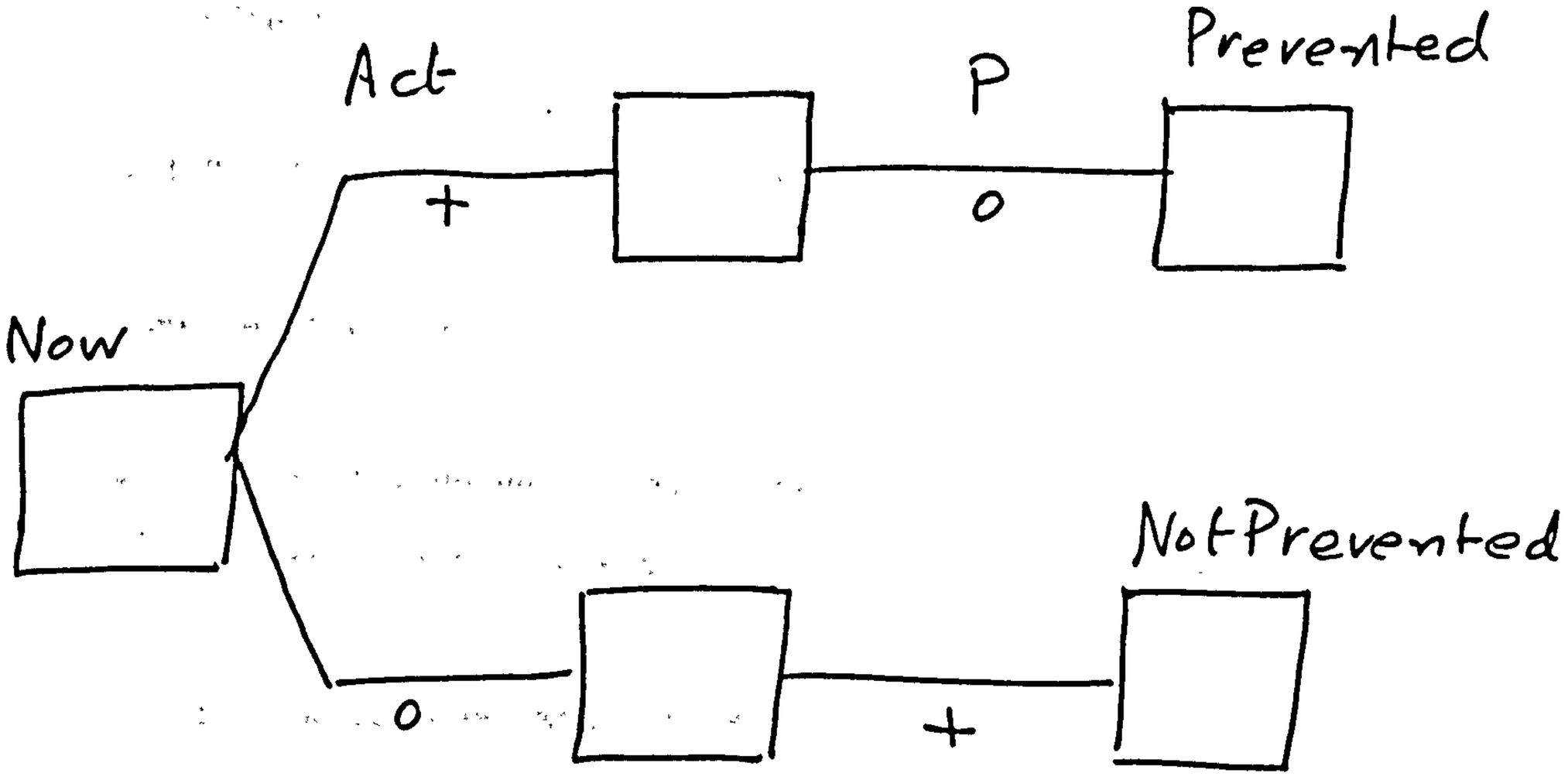
Under/2



under/3



Prev/5



good effect, proved by proof P, then NotPrevented is better than Prevented, and so choosing to act is unwise. And conversely if E is a bad effect.

4.4.2. Effects on future actions

Consider the example 4 above.

4 A: I'm going to have some of that crumble.

B: I wanted to give it to Elaine and Richard.

Clearly what one wants to say about this is that the result of A's eating the crumble is that it will undermine B's plan to give it to Elaine and Richard. There will be no crumble to give. But the trouble about seeing the result of A's action is that the good A seeks, that Elaine and Richard have crumble, is provable neither before A's action, nor afterwards. So how can their not having it be a result of what A has done?

The SOA diagram explains it. This is an occasion when one has to look further into the future, to after when B has acted. See Prev/6. i shows no effect. ii does. As an interaction of proofs, this can be drawn as Prev/7. The node in contention will be something like "crumble exists".

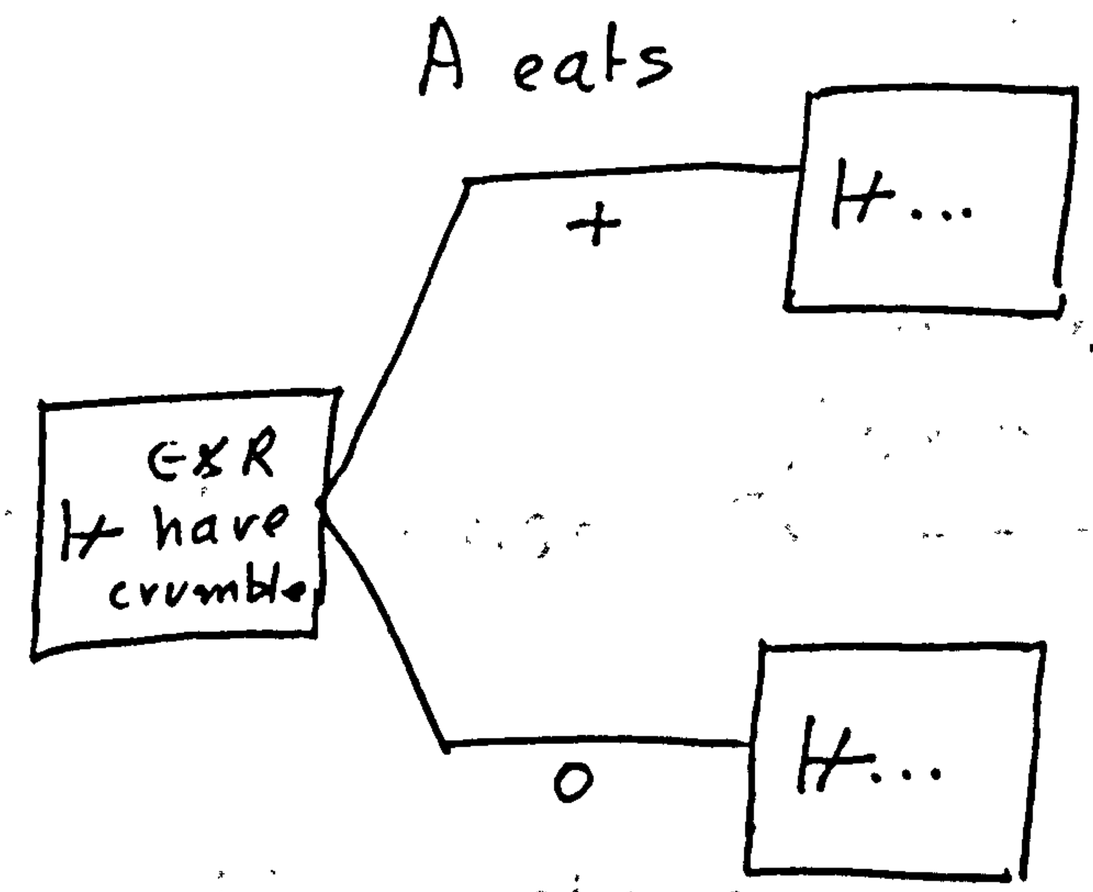
Defining the results of actions by looking at proof trees sound only in the future produces no novelty.

4.5. Changing the apparent merits of a plan

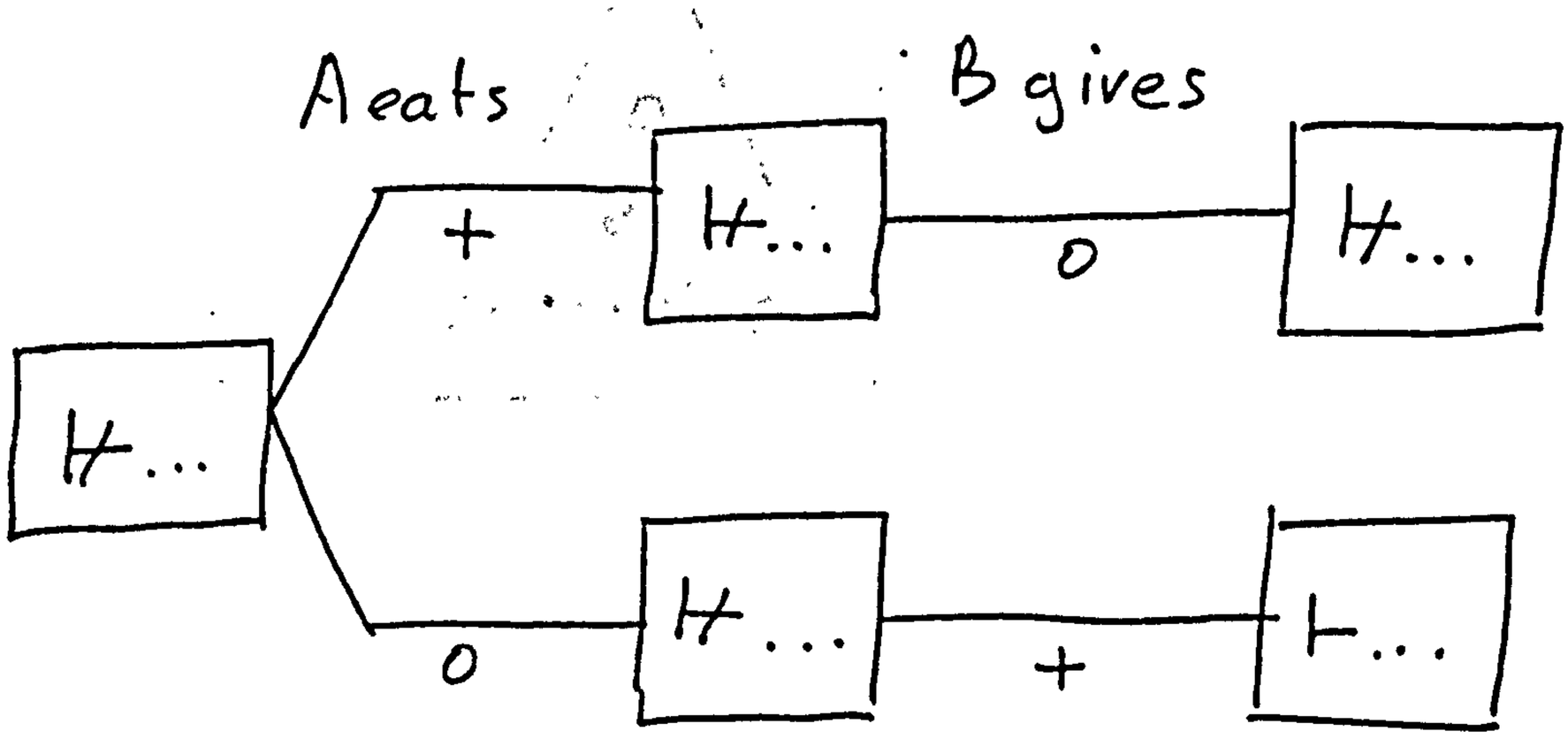
If you are going to perform an action, it must be possible and worthwhile. It is possible, "sound", if its preconditions are true,

Prev/6

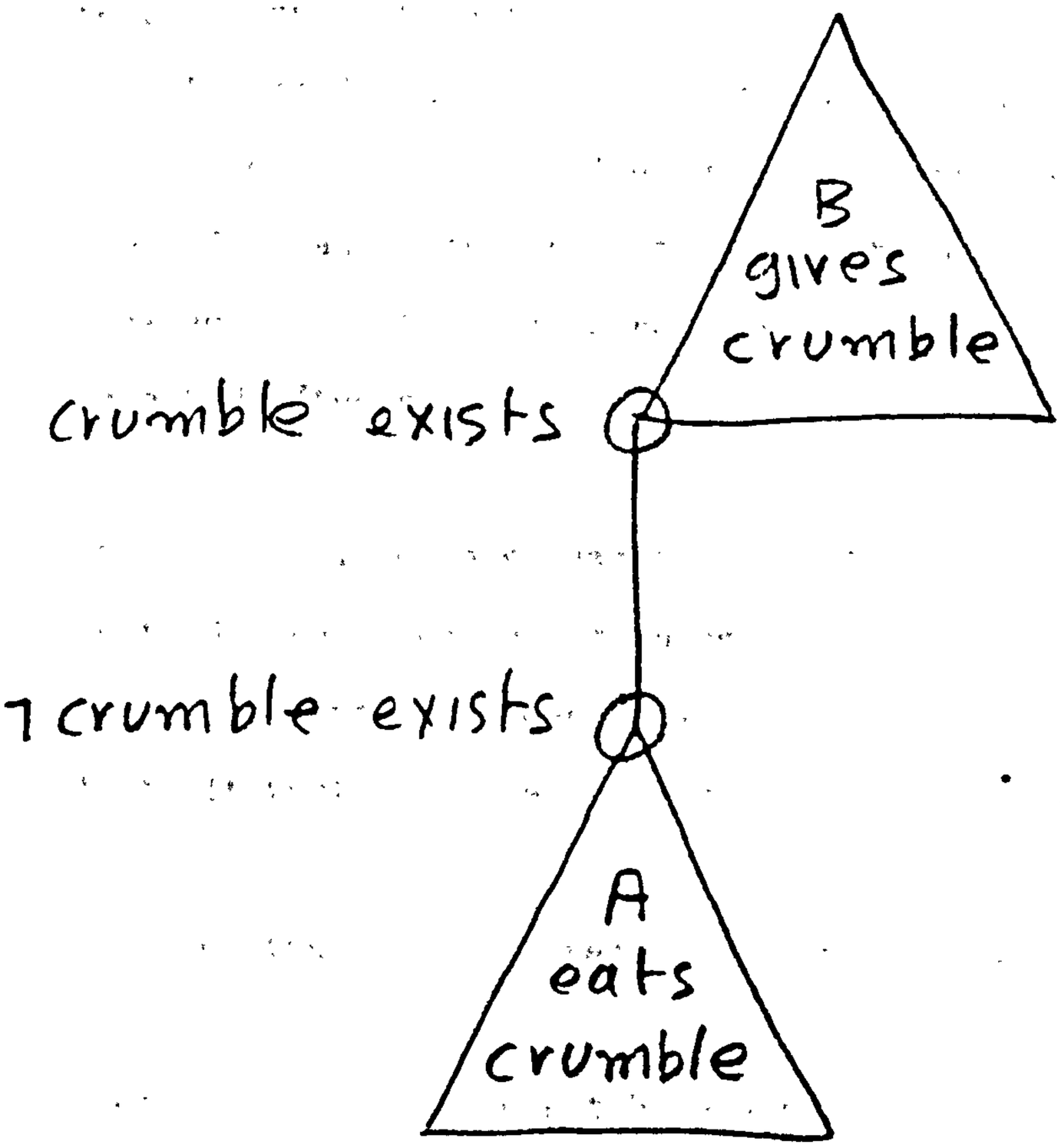
i



ii



prev/7


B
gives
crumble
crumble exists
¬ crumble exists
A
eats
crumble

so that when the actions in it are performed, it will run to completion. It is worthwhile if its good results outweigh its bad results and the effort involved. (I use "result" to be more general than "effect", and to include what an act prevents.) I do not want to get involved in weighing good results against bad, so I shall simplify, and say an action must be needed and safe. An action is "needed" if it has good results. An action is "safe" if it does not have bad effects.

If an utterance does change your estimate of the action's merits, it must do so because it persuades you differently about either its soundness, or its worth. An action's soundness is just a matter of fact. It worth is more complex. The utterance must persuade you

- either that it has a result that you thought it didn't.

- or else that it doesn't have a result that you thought it did.

The results that matter are that a valued fact V gets added to a SOA or deleted from it. If this happens it must be because a proof tree P which has the consequence V goes from sound to unsound or vice versa. If V is wanted, these are the transitions that count as a result.

A fact may be added

P sound after	<-->	- P sound after
- P sound before		- P sound before
V wanted		V wanted
P sound after	<-->	P sound after
- P sound before		P sound before
V wanted		V wanted
P sound after	<-->	P sound after
- P sound before		- P sound before
V wanted		- V wanted

A fact may be deleted

- P sound after P sound before V wanted	<-->	- P sound after - P sound before V wanted
- P sound after P sound before V wanted	<-->	P sound after P sound before V wanted
- P sound after P sound before V wanted	<-->	- P sound after P sound before - V wanted

Changes in your estimate of a result such as

- P sound after P sound before V wanted	<-->	- P sound after - P sound before V wanted

are not important, because neither before nor after the change is V a result.

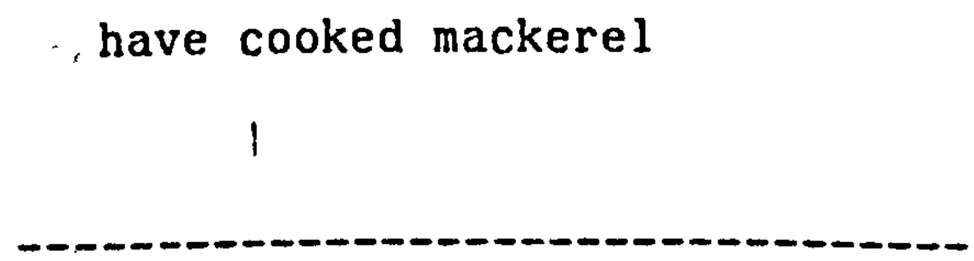
If "V feared" is equivalent to "-V wanted", these tables cover bad results too.

4.6. Unexpressed plans

So far I've talked about how utterances can attack or support plans that have been announced or suggested by their agent. But there are some utterances whose point you can only see if you are prepared to see them as affecting unannounced plans. For example

Sp: I've bought some mackerel

points out that a plan like this



.. various cooking actions ...

|

have mackerel

is, surprisingly, sound. Sp presupposes that eating mackerel is good and not just at the minute possible, so that his utterance points out novel good effect. Similarly

Sp: The bathroom shelf is about to collapse

suggests that a plan to buy screws and fix it is, surprisingly, needed.

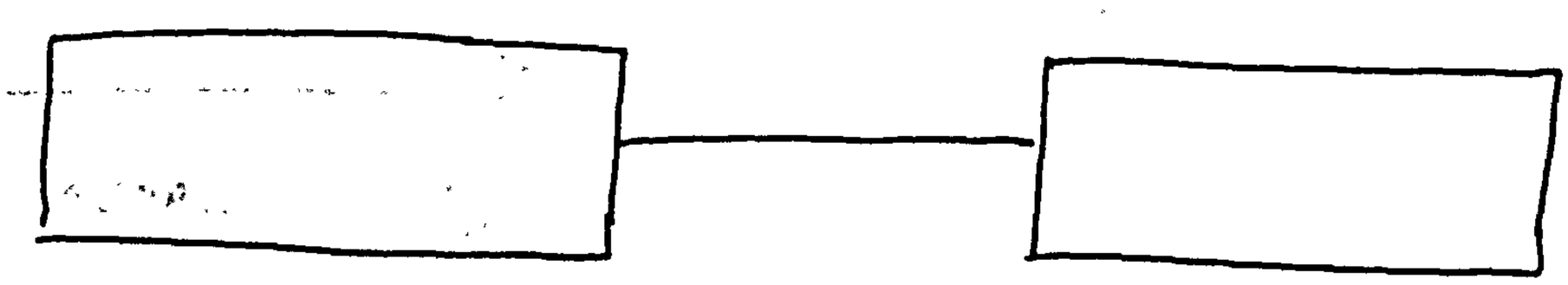
One way of explaining the point of such remarks would be this. Initially Hr expects the world to carry on just as it is. Suddenly, (by some undescribed process), he realizes both that he has options, and that one choice is better than the other. His expectations change from Unsaid/1i to ii. But doesn't this deny my claim that one chooses an action because of one's changed estimate of its effects? For where in Unsaid/1i is the action "cook mackerel" that Hr starts by choosing not to do? Nowhere. How can a remark change Hr's estimate of an action he doesn't see as an option?

This is a good processing point, but a poor logical one. The logical come-back is that in fact one is always contemplating all possible actions; or, even if one isn't thinking about them, they are laid out before one in some Platonic sense, visible if one cared to look. Then Hr's real expectations are as in Unsaid/2. Before Sp speaks, they are as in i. No advantage comes from cooking the mackerel. But after Sp speaks, advantage is possible, and Hr should choose accordingly.

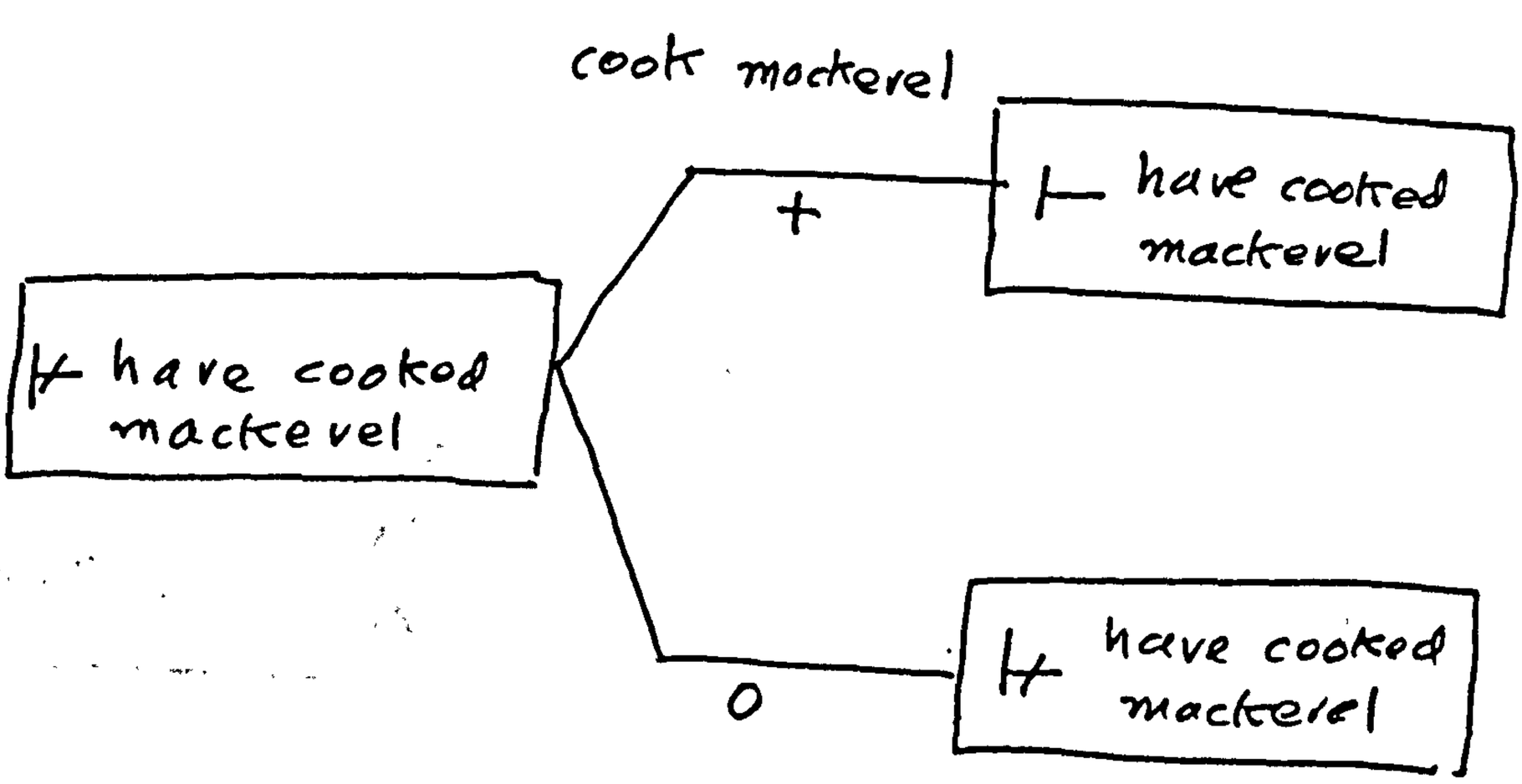
Nevertheless, this alone will not do from a processing point of view.

Unsaid/1

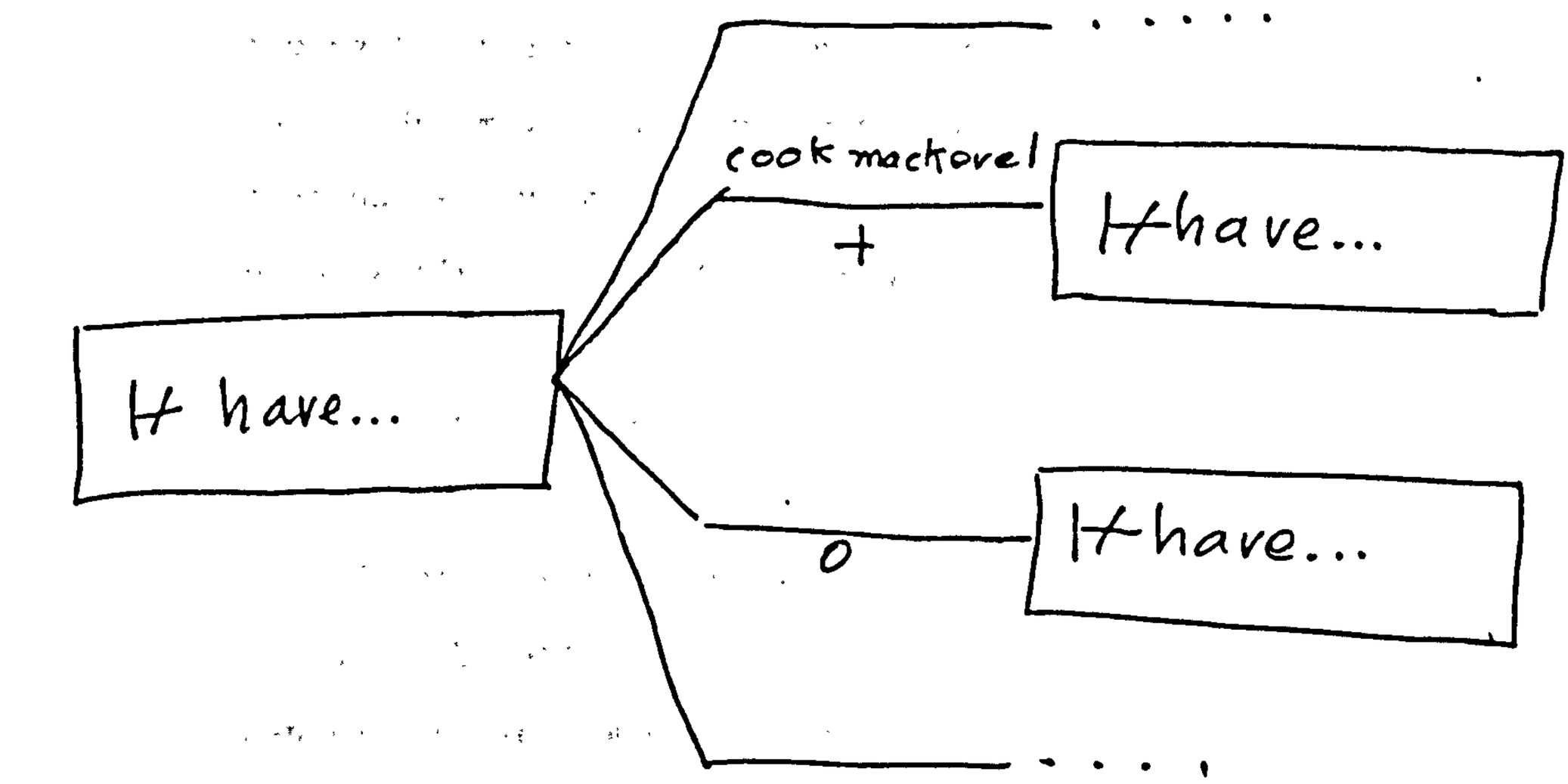
i

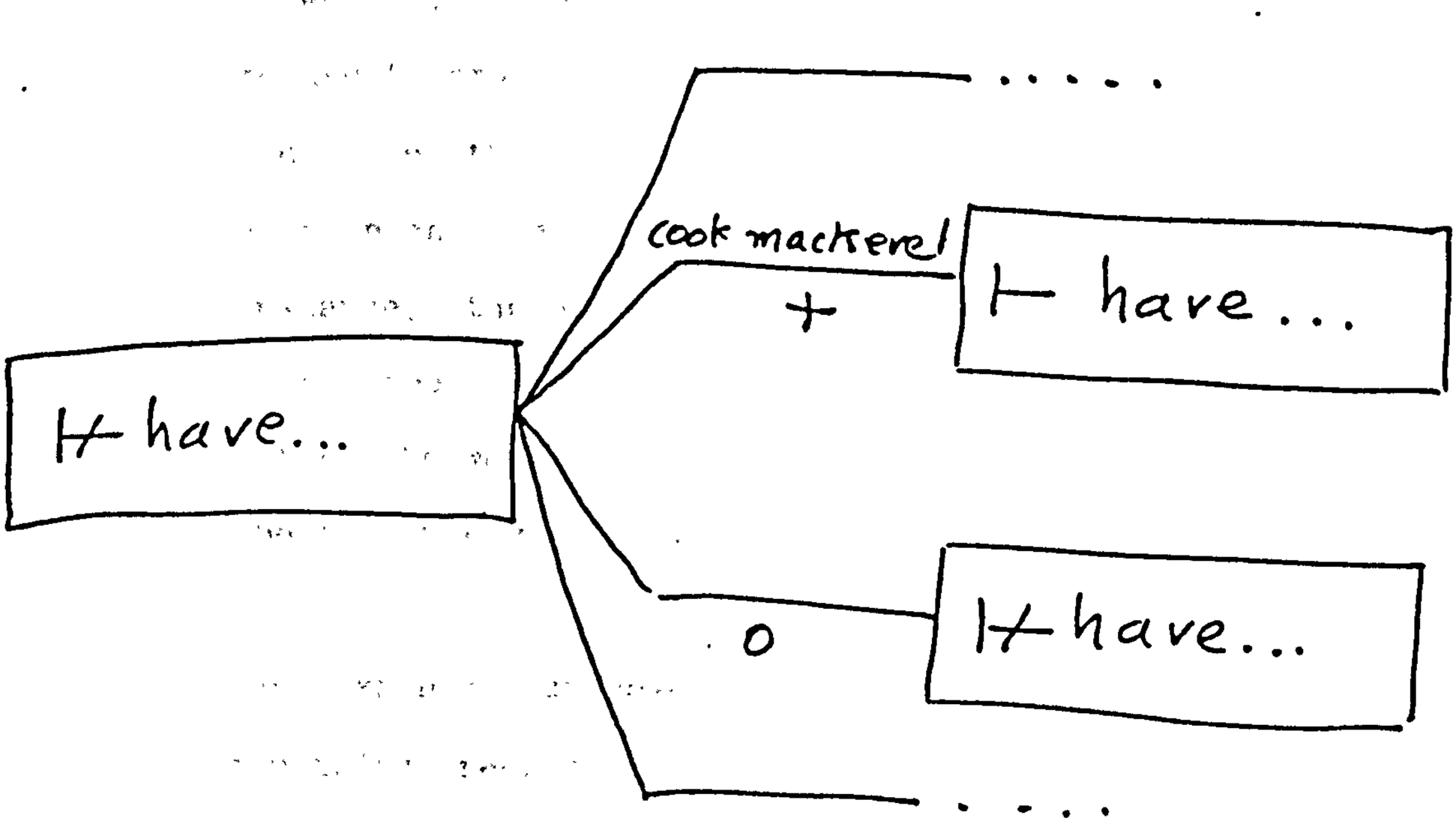


ii


cook mackerel
have cooked mackerel
+
have cooked mackerel
0
have cooked mackerel

Unsaid /2


cook mackerel
+
H have...
H have...
o
H have...


cook mackerel
+
H have...
H have...
o
H have...

When Sp speaks, Hr is not in fact contemplating any action, and he can't search forward from it. If he wants to find what action Sp is trying to affect, he is going to have to guess it merely from what Sp said. He may though be helped by reflecting that it is more likely that Sp is trying to persuade him to an action he has not chosen than pointlessly to dissuade him from one he has not chosen.

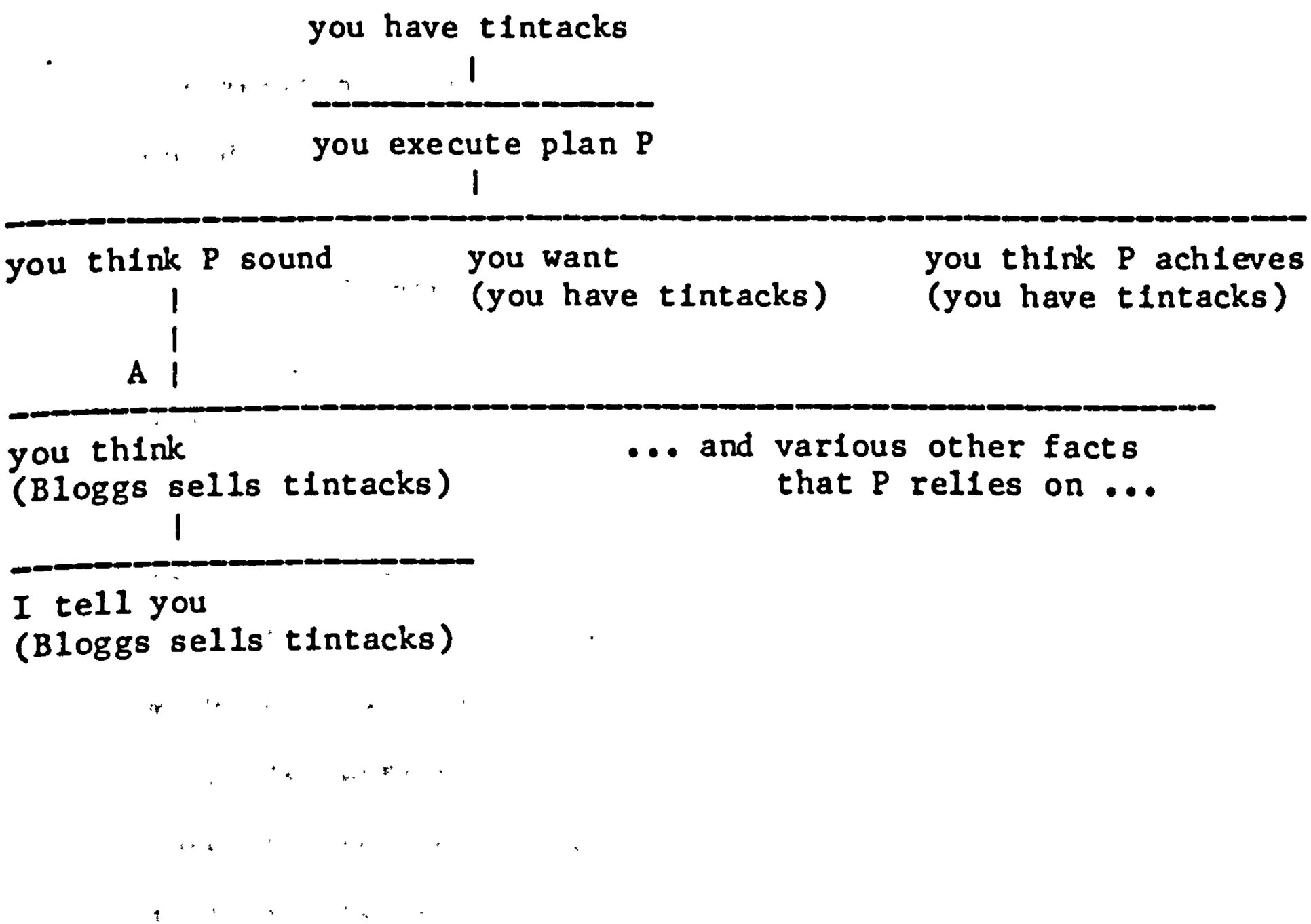
4.7. A wrong approach to finding Sp's benefit.

I did at one time try a different framework for demonstrating Sp's ultimate benefit from speaking, which I ended by rejecting. The central notion was to construct a plan that could be attributed to Sp, which included his utterance as a sub-action and which had some desired fact about the world as the goal. All other actions, especially the utterance, would be explained as rational because they tended to produce that goal. What would be distinctive about the plan would be that it would contain inferences which allowed predicates over plans, as well as over things in the material world. For instance, the plan could involve rules that allowed the inference of facts about the world from facts about the execution of plans to alter the world, and of facts about knowledge about plans from facts about knowledge of the world.

For example, suppose you were planning to buy tintacks. Buying something requires knowing where they are sold. If you do not know this, I can help you by telling you that they are sold at Blogg's. I might explain my saying this as part of a plan of mine like the one in Failnote/1. An example of a rule that entails facts about plans from facts about knowledge of the world would be the rule that allows step A. It would be an instance of a rule schema such as

A thinks plan $\bar{P}$ is effective if

Failnote/1


you have tintacks
you execute plan P
you think P sound
you want
(you have tintacks)
you think P achieves
(you have tintacks)
A
you think
(Bloggs sells tintacks)
... and various other facts
that P relies on ...
I tell you
(Bloggs sells tintacks)

P has preconditions C1...Cn &

A believes C1 & ... & Cn

The merit of this approach is that it fits my intuitions about what I would say if asked to explain why I had told you that Bloggs sold tintacks, at least when handling utterances that help someone get a single good thing. But there are two problems with it.

4.7.1. It can't see that bad is better than worse

It won't explain my telling you something that makes you give up a plan that you thought was a good one. Suppose you are going to buy milk. I tell you that the shops are shut. You no longer think the plan is effective and so you give up the plan. Clearly there is a benefit. You do not waste effort. But that is not the right sort of fact to be the goal of a plan like Failnote/1. The goal has to be a desirable fact about the world. But what can that be? That you do not have milk? That is hardly a benefit, and anyhow could hardly be the goal of a plan. It is true already.

4.7.2. It can't see that good is worse than better

Suppose you and I are normally avaricious persons. There are two sums of money, $10 and $1000. We will each get just one of them. You do not know it, but there are two course of action open to you, one to get each of the sums of money. Disingenuously I tell you things that lead to you to choose the course of action that leads to you getting the $10 while I get the $1000. It will be possible to give an explanation of what I said in terms of my ensuring that you get $10. This is wrong. The explanation should be in terms of me ensuring that you don't get the $1000. But since your having $10 is an absolute good, my action is explained.

In both of these, what is missing is an ability to contrast one outcome with another. In the first, a bad state (not having milk) is still better than the alternative (not having milk and having gone to to shop to get it). In the second, a good state (having $10) is still worse than a the alternative (having $1000).

Any method that seeks to explain an action by saying that it leads to a good state, rather than the best state, will have these problems.

4.8. The arguments for and against an action

To summarize: there are four sorts of argument for or against an action. They all turn on what happens to some valued fact V. Here is a table that shows how these cases affect the merit of the action.

V	V before action	V after action	Merit of action
good	unprovable	provable	good
bad	unprovable	provable	bad
good	provable	unprovable	bad
bad	provable	unprovable	good

But finding which facts are provable and which are unprovable in a SOA is in general too difficult. It is better to look at particular proofs of V, and replace the notion of "provable" with that of "having a sound proof". If P is a proof of V, the table looks like

V	P before action	P after action	Merit of action
good	unsound	sound	good
bad	unsound	sound	bad
good	sound	unsound	bad
bad	sound	unsound	good

Attempts to change one's estimate of the merits of an action are attempts to show that some proof P of some fact V in fact do or in fact don't fit one of these patterns.

4.9. Conclusion

Plans may or may not seem to their planner to be worth doing. If a speaker gives the planner new information, it may change the planner's estimate of the worth of the plan. This change in estimate may itself be beneficial to the speaker:

- either because a friend is provided with a new plan whose execution will please the friend more than the old plan's execution would have
- or because the execution of the changed plan will please the speaker more than execution of the unchanged plan would have, either because of its immediate effects, or because it facilitates another of the speaker's plans.

Chapter 5. Indirect Communication

5.1. Introduction

Consider these examples.

A: Do you like Florence?

B: We've gone there three times.

A: Have you emptied the bins?

B: I'm just about to.

A: Is junction 30 open?

B: I came through it on the way here.

What happened in them? B has given an answer to the question that A asked. He has done it, not explicitly, but by saying something about his plans from which the answer follows. In this chapter I want to suggest how this information is derived, and that that information can be used in the same way as information from an explicit utterance.

In outline: if B knows A has a plan, then B must think that A has certain beliefs and values. He must think that A wants the effects of the plan, believes that the plan is sound, and so forth. If the plans that B thinks that A has have changed, then the beliefs and goals that B must believe A has must have changed too. If B takes A to be authoritative on the subjects of these beliefs, B's beliefs will change too. Now A can exploit this. He can change the plans that he is known to have, intending that this shall change B's beliefs. The change produced will then act just as a change produced by utterance

would have. This change is indirect communication. The ultimate goal of the utterance will be an effect of that communication.

Here are some examples of bits of dialogue. They are collected because in them one of the speakers is expressing a belief or want, without however overtly saying what it is he wants or believes.

Expressions of fact

A: I'll wear my new jersey.

B; I'm just going to put the sleeves in.

B is saying "... and I would only put them in if they weren't already there. So you can't use your jersey".

A: I want to use the car.

B: I was going to take it to the garage.

Again, B is saying "... and I would only do that if it wasn't running properly. So ... "

Expressions of wants

A: Pass me the sponge.

B: I'll wipe the table.

A is saying "I will only want the sponge if I am going to wipe the table. So I want the table wiped. Will you do it?". B agrees. If you find this counter-intuitive, try reading A's remark in an impatient tone. Such exchanges really occur.

A: My cheque book isnt where I thought it was.

B: It's by the cooker.

A is saying "... and not knowing where it is affects my plan. It would do this if my plan depended on knowing where it was. So I want to know where it is".

A: The 5:14 bus has already gone.

B: I can give you a lift.

A is saying "... and my plan relied on catching it. Since my plan has collapsed, its ultimate goal of going somewhere is still outstanding. Can you help?"

The two main questions are:

What facts does a person having and changing a plan entail?

How are the plans that a person is known to have changed?

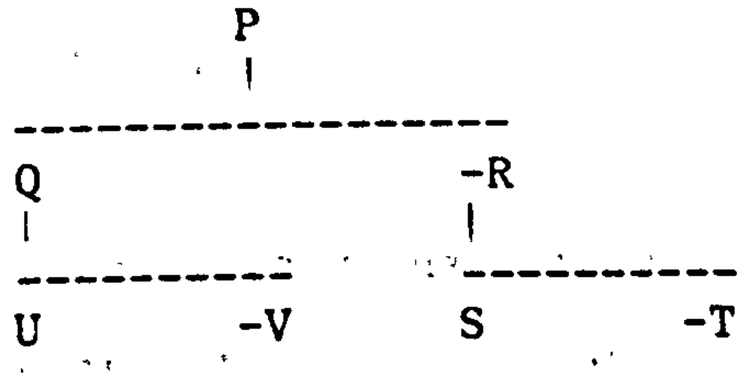
5.2. What does a plan entail?

5.2.1. Facts about beliefs entailed by a plan

I have claimed a plan can be seen as a proof tree proving a goal. If a person adopts a proof tree as a plan he must believe it will have a good effect and will lack bad effects. A fact is an effect of a set of actions if it occurs in no sound proof tree which includes none of the actions, but occurs in a sound proof tree which does include them. The proof tree for the goal of the plan is the plan itself.

The proof tree can be divided into those nodes that are expected to be true after execution, and those that are beleived to be true before. Those nodes expected to be true efter execution are those

that lie above any action. Eg in



P,Q,-R are false now and true later; U,V,S,-T are true now. So the holder of a plan will believe that:

> The nodes at its fringe that are facts are true, and those that are actions are possible; otherwise his actions wouldn't have their intended effect.
>
> The nodes above its fringe, that rely on nodes that are actions, must be false; Otherwise they would not be effects of the plan, and would have been assumed, instead of being planned to.

Analogously, he must believe that any proof tree that might be an effect of his plan and that would then prove some feared fact does not satisfy both those conditions. Either some of the nodes at the fringe of the tree are false, or the nodes above the actions in the proof tree are true anyhow. In an example such as

> A: Is this water safe to drink?
>
> B: I drink it.

what agitates A is that he doesn't know whether the water is contaminated, and so he doesn't know whether a certain bad side-effect will in fact occur.

```
          - A thirsty                          A ill
                |                               |
--------------------------------  --------------------------------
A has water                     A drink water          water contaminated
```

When B says he drinks it, it must mean he believes that his plan to drink it has no bad effect. This can only be true if the water is not contaminated. So he indirectly communicates that he believes it isn't contaminated, and so answers A's question.

In this example, the fact that is indirectly denied is obvious. But it need not be. To take an abstract example, suppose A wants to do this

```
    G
    |
----------
H       - Q
```

but is worried about a side-effect like this (B is something bad).

```
    G             B
    |             |
--------  --------------------
H     - Q              J     K
```

Now a friend tells him that his plan is in fact safe. A can go ahead. He can infer that -(J & K). But he does not know which is false. He needs some other test. It may be that he knows that his friend thinks J is true, so it must be K that is false. But if he can't reason like that, then he may be able to say that there is only one fact whose falsity would have an effect on his own plan. So if he believes his friend is trying to affect his plan, it must be that fact that hs intends to deny.

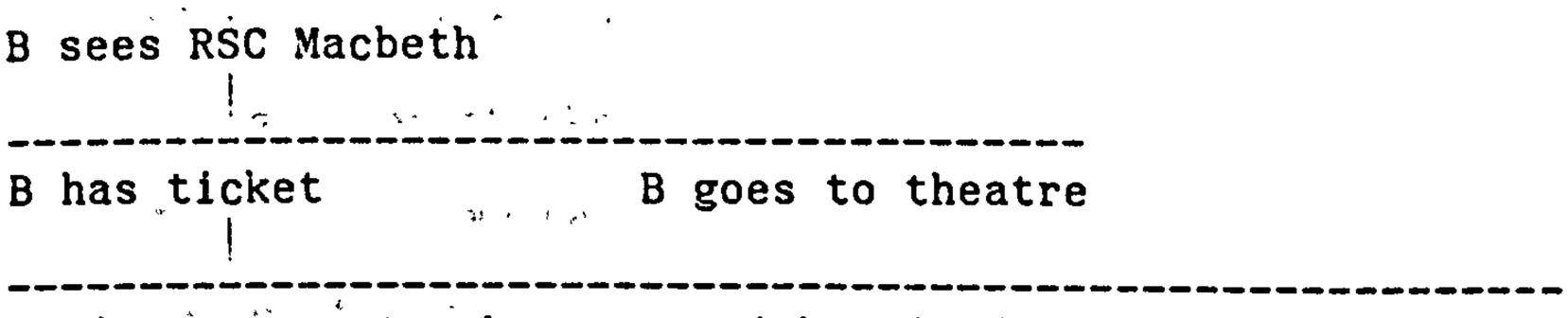
That a person abstains from an action may also be informative. If I say that I won't do such and such, you may conclude that it is because that action wouldn't have a good effect or would have a bad one.

Here is an example which illustrates both those points: an abstention is informative, even though the hearer has to make a choice about which of several facts that might have been denied actually has been.

A: I'm going to try and see the RSC Macbeth.

B: I didn't even try to get tickets.

Presumably B's plan is



Several things could go wrong with this plan. But the only one of them that could also affect A's plan, which is presumaby very similar, is that there might be no tickets for sale. The main good effect of B's plan will be unsound. And that is presumably what B is trying to convey.

5.2.2. Facts about values entailed by a plan

A plan entails facts about its holder's goals as well as about his beliefs. Suppose we know a plan of his that he takes to be sound, but which he hasn't yet started to execute. Given the proof tree of a good effect, he will fear anything that makes it unsound. If it is unsound, he will want anything that makes it sound. Analogously with a bad effect: he will want anything that makes it unsound, and fear anything that makes it sound.

In fact he may desire facts that do less than turn a good effect from being unsound into being completely sound.

Suppose a good effect is already sound. Suppose some of the facts that the planner thought he had to achieve by himself become true without effort on his part; perhaps they are done by someone else. The effect is still sound, but it is better for the planner.

Suppose a good effect is unsound. Suppose some of the facts that the planner hasn't achieved become true. But the planner may think he can achieve the remaining facts in the fringe, or thinks that they may become true, perhaps by the action of someone else. The plan is still unsound, but it is better for the planner.

It should also be possible to infer that an agent wants the topmost fact that is brought about by his plan. But there is a catch here. Suppose you act so as to achieve something that I find it hard to believe that you might have as an ultimate goal. I may in fact know all your plan. But if I don't know that it's all your plan, I may be more likely to infer that it is part of a plan to some more credible goal, rather than that I am wrong about your goals. I remember a "Peanuts" cartoon in which Lucy has this problem.

Lucy sees Schroeder with a record.

Lucy: What are you going to do with it?

Schroeder: Listen to it.

Lucy: Are you going to sing Along wit it? Or dance to it?

Schroeder: No. I'm just going to listen to it.

Lucy: I never heard anything so dumb.

5.2.3. How does someone change the plans that he is known to have

The change in a person's plans that matters in indirect communication is not how his plans alter, but how what is known of his plans alters. Altering a known plan (whether it is made better or worse or just different) and letting it be known that one has a plan where none was known before are for these purposes the same.

How can he alter what is known of his plans by what he says? It seems there are three ways. I shall later try to show how part of this range can be seen as special cases of a more general process.

5.2.3.1. Description of his plan

The planner can just announce what he is doing. He says "I am going to the shops" or "I shall buy the vegetables tomorrow".

5.2.3.2. Showing that he has grounds for making a plan

He is thought to have a plan P (perhaps the null plan). He says something about what he wants or believes. If he really wants or believes that, and if he is a rational planner, then he must have a different plan P'. The contrast between P and P' is where the indirect information comes from. This indirect information leads to a benefit for Sp.

5.2.3.3. Inviting the recognition of the plan that he has

Or he is thought to have a plan P (perhaps the null plan). He says something that is likely to get a response from B. This response will only help A if his plan is in fact in some way different from how B initially thought it was - if it is P'. The contrast between P and P' is where the indirect information comes from. This indirect

information leads to a benefit for Sp.

Here are some examples where B guesses what A's plan is and responds appropriately.

A: Do you mind if I turn the fire off?

B: It's on a time clock.

A: Have you been to the bank recently?

B: Yes. How much would you like?

But this scheme does not handle examples like this one.

A and B are both about to go home at the end of the day.

A: The 5:14 bus has gone.

B: Can I give you a lift?

A has said something that can count as an indirect request. At any rate, he has indicated that he has an unsound plan. He wants to get home by catching the bus; but a precondition of that plan has failed. So he wants to get home some other way: perhaps by getting a lift from B.

But how can he expect B to see this? The obvious thing to say is that this exchange fits scheme 2. He has a plan. He discovers new facts. So his plan is changed. The contrast of his former and latter plans reveals some of his goals to Hr. The objection is that initially B did not know that A had a plan. B has to hypothesize A's plan (to go home by bus) before he can see the effect on it of A's remark.

When what A says is a question, A can expect B to hypothesize A's

plan by straight-forward plan recognition. A's question has an effect. What goal could that question serve? One that the answer could affect. What plans could it affect? One that relied on the answer: for instance ... And hence A's plan has been recognized. But the normal response to A's remark is not quite a reasonable goal. But if B looks for indirect communication derived from contrast of his former and current picture of A's plan, he can see Sp's goal.

When what A says is a statement, the straight-forward plan recognition can't work. A's remark is not an actual part of some plan of his. So A can't expect B to find his plan, and then, after having done that, to derive indirect information from a contrast. A has to expect that B is going to look for indirect information in its own right, and that he will hypothesize plans just because they have the property of being affected by what Sp says. But is he licensed to expect this?

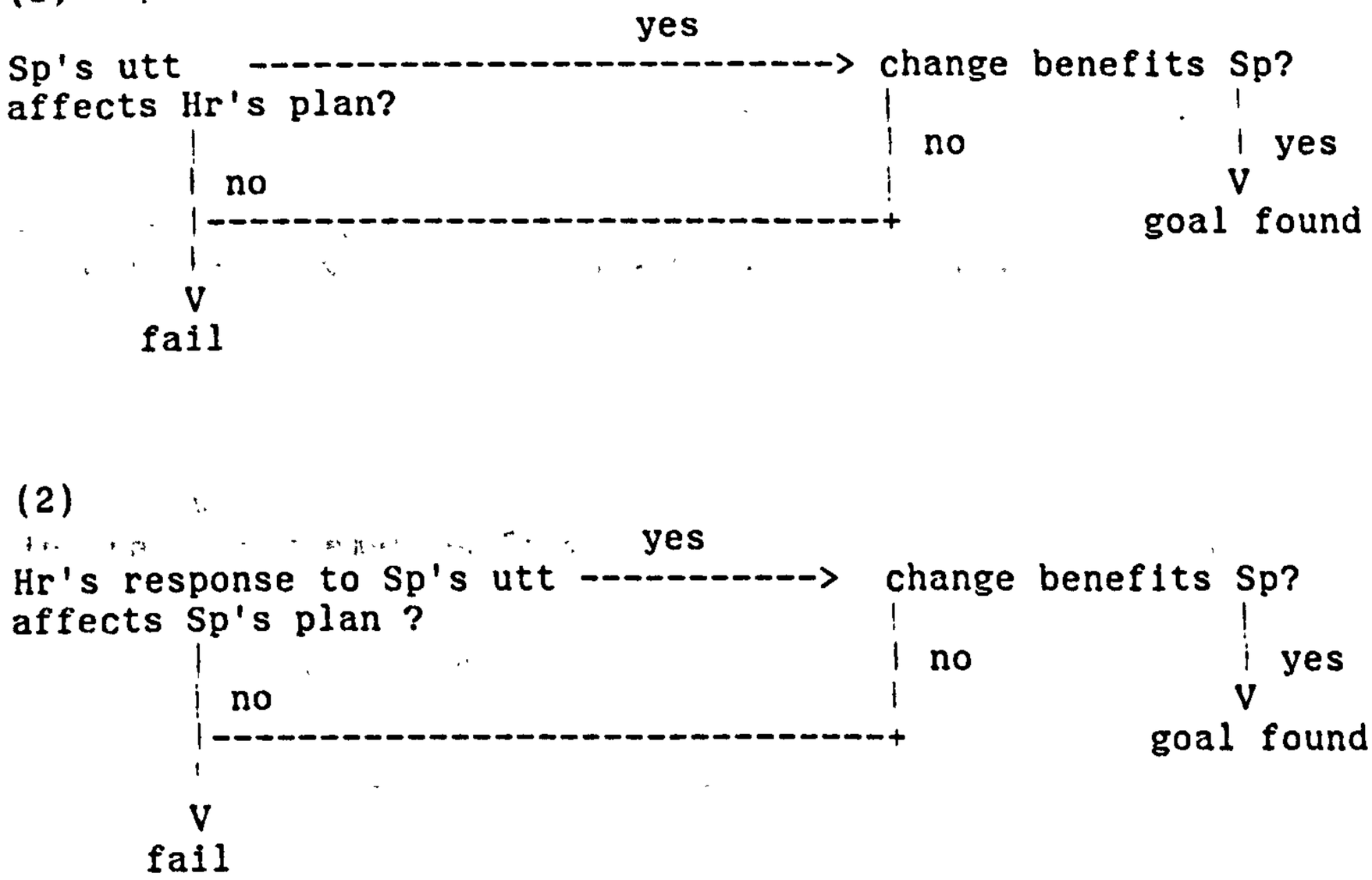
I could propose that, when people listen, they do expect to have to look for indirect information. This would be a new principle governing what they do. But I would rather suggest that there was one scheme for looking for the benefit of Sp's remarks of which this was a special case. This scheme is known to be practised by all hearers, so all speakers can rely on their hearers getting the conclusions that drop out of it.

What a speaker expects his hearer to do with his remarks to see the point of them can be described with a flow-chart. Take this flow-chart not as a claim about the order of events but the conditions on saying that something is the intended benefit of an utterance.

The simplest cases of benefit are (1) when Sp affects Hr's plan to Sp's benefit, or (2) when Sp asks a question, and Hr's response to it

will affect Sp's plan to Sp's benefit.

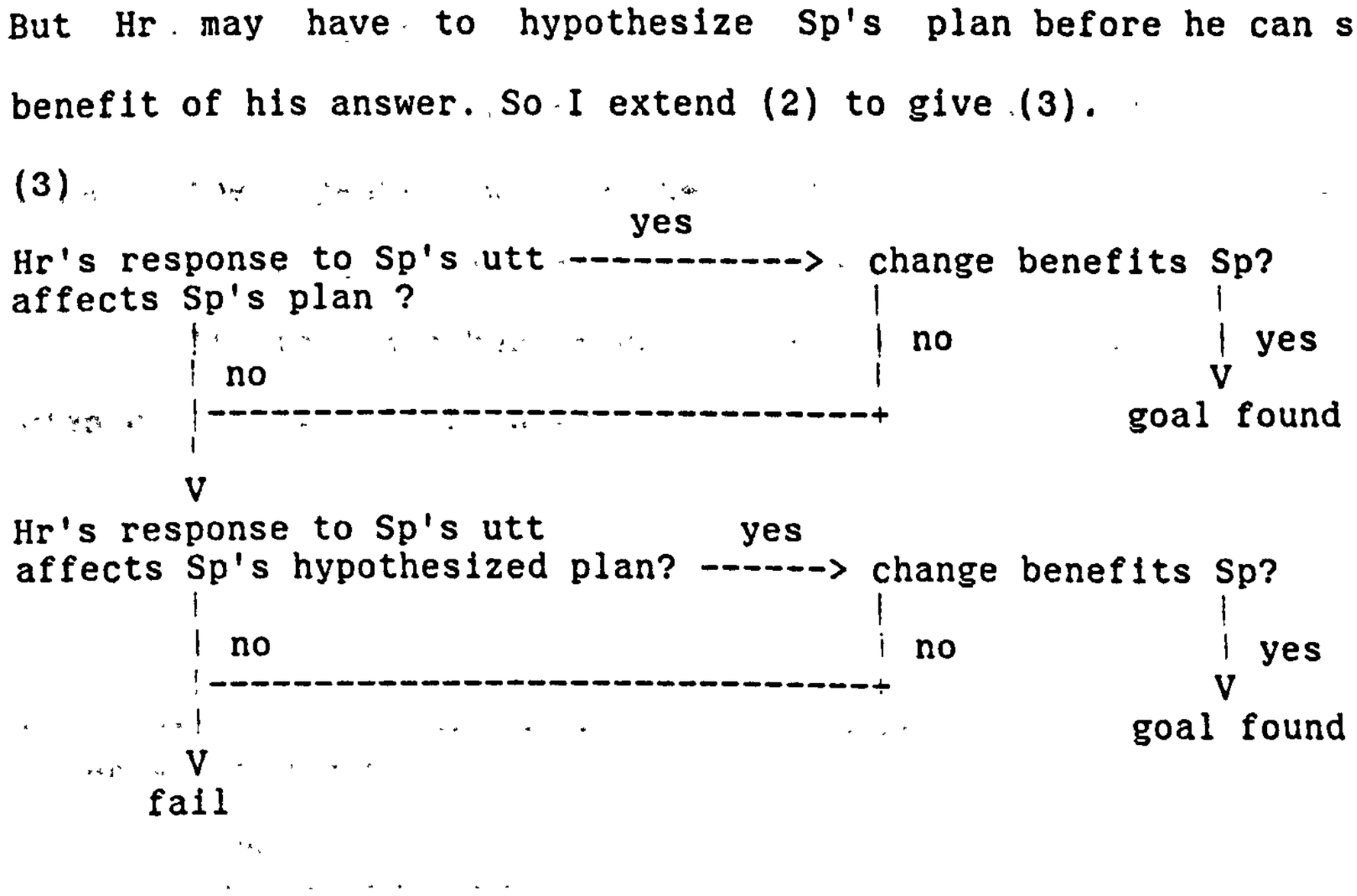
(1)



(2)

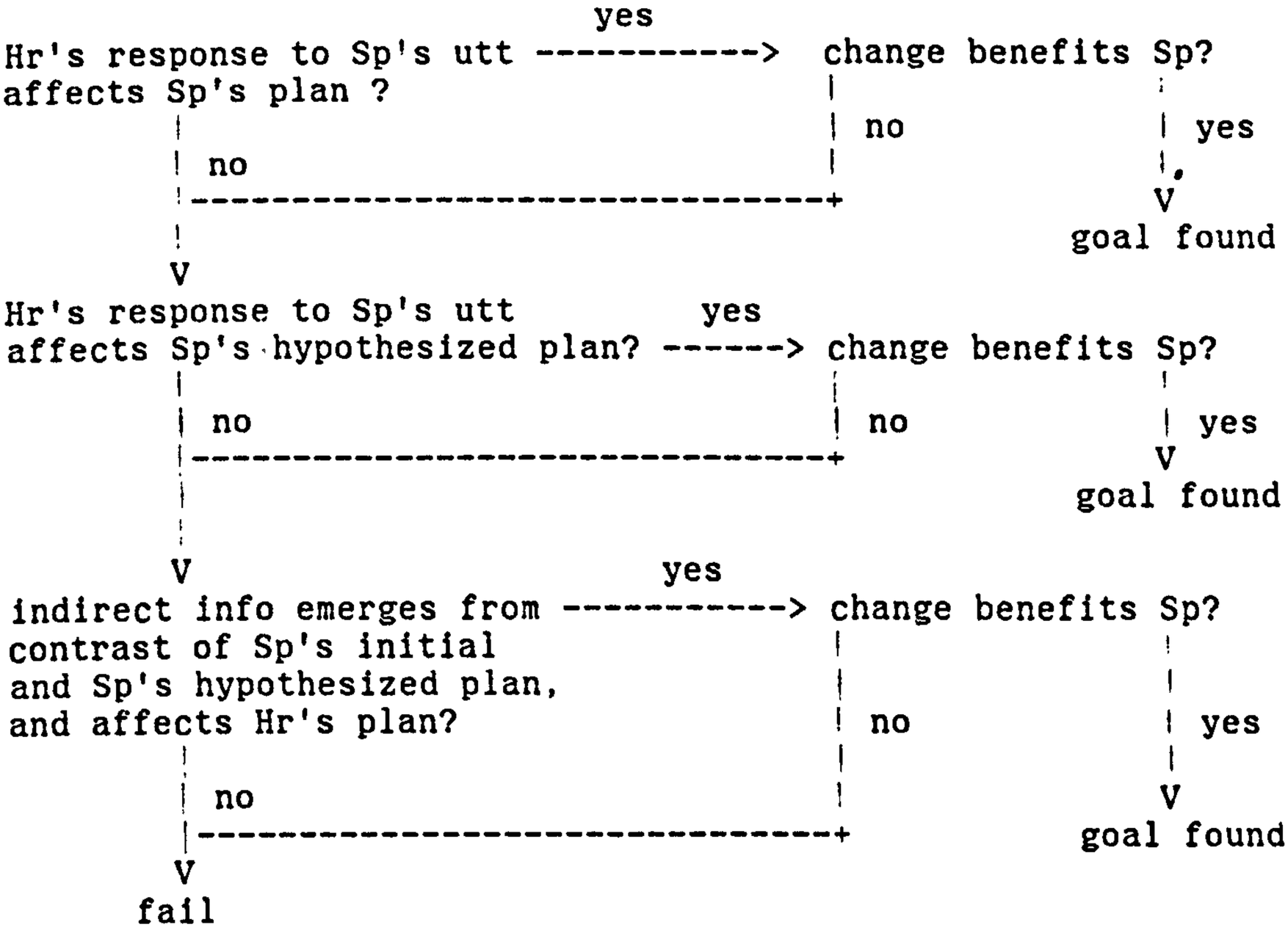
But Hr may have to hypothesize Sp's plan before he can see the benefit of his answer. So I extend (2) to give (3).

(3)



And last of all I want to allow for the case when the indirect information from the plan that Hr has to hypothesize about Sp is what matters.

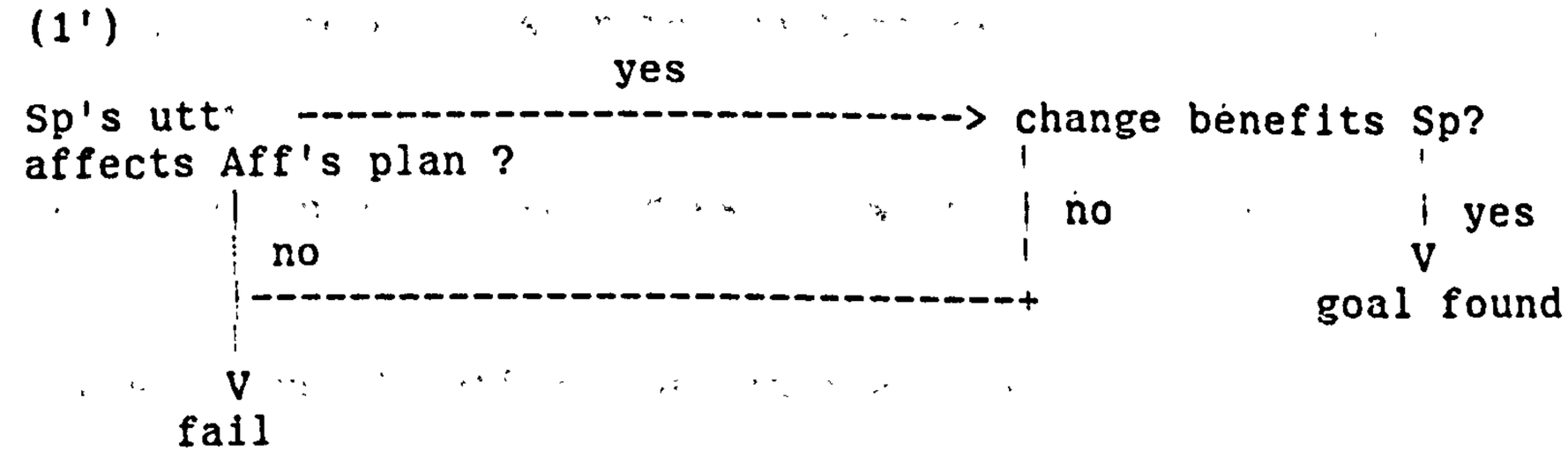
(4)



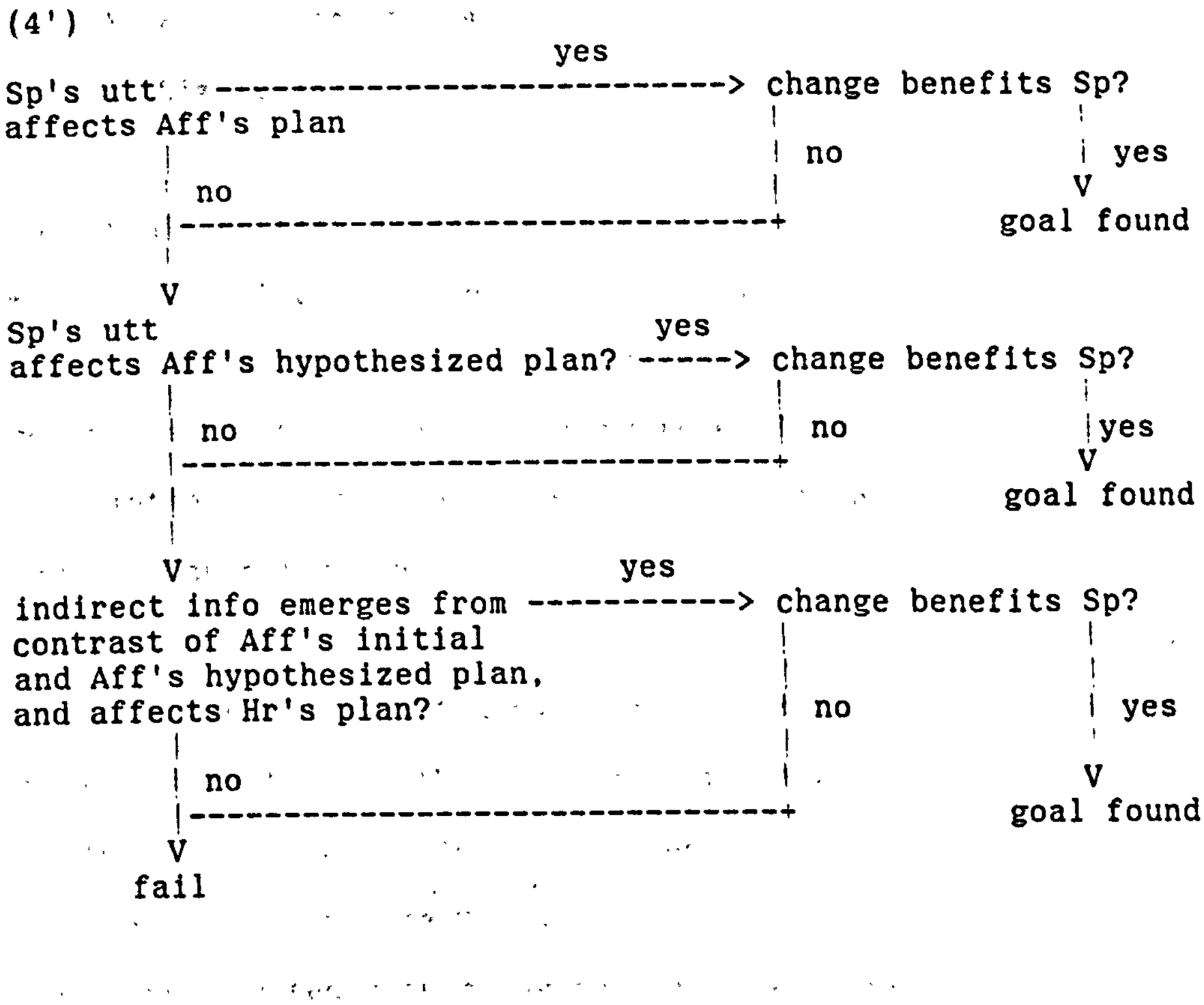
Now making a statement and asking a question are in this way alike. There is a person who precipitates an information transfer. This is always the speaker, whether he is stating or asking. And there is a person whose plan is affected by that information. I'll call him "Aff". If the speaker made a statement, Aff is the Hr (at first glance, anyhow). If he asked a question, Aff is the Sp.

Now I can collapse (1) and (2) into one chart, (1').

(1')



What happens if I relabel the roles in (4) in the same way? After all, (2) is just a part of (4).



Now I claim that this is the scheme that all Hr's apply to find the goal of Sp's utterance. (1), (2), (3), (4), are all special cases of it, perhaps with some parts omitted, perhaps with a particular choice made between Sp and Hr for the role of Aff. That choice has to be compatible with the choice about whether Sp made a statement or asked a question.

But I want to say that Sp may be attempting to alter a plan of his own that he expects Hr to hypothesize. That is the extension I have to make to cover eg the missed bus example. So I have to allow Aff to be either Sp or Hr, regardless of whether Sp was stating or asking.

Another justification for making the important distinction that between Sp and Aff, rather than that between Sp and Hr is that Sp may be able to achieve indirect communication by saying things that alter Hr's estimate, not of Sp's plans, but of some third party's plans. Eg

A: Have we got the salad?

B: Liz is just bringing it.

And many charity advertisements work similarly. "X is without clean water" is the pointing out of the falsity of a thing good for X. The advertiser hopes his reader is altruistic towards X, and will act to serve the revealed goal. The structure of this intention is just like the intention of the husband (with a certain sort of marriage) who says "I haven't got a clean shirt".

One question I can't answer is, how is Hr to hypothesize the right plan? In outline, the heuristic must be, not: is this a plan to a credible goal that could be advanced by the information transfer that Sp's remark has started?, but: is this a plan to a credible goal that could be affected either way by that transfer?

5.3. Conclusion

If a planner is known to have a plan, he is also known to have certain beliefs and values. If this plan changes, the beliefs and values he is supposed to have must change too. The planner may be able to say things which imply that his real plan is different from the one he is thought to have. Thus it becomes known that he has different beliefs and values. These may influence other people's plans just as if they had been said. So when looking for the benefit of an utterance, one has to see if it has provoked any of this sort of indirect communication, and if it has, one has to see whether that is how the benefit of the utterance accrues.

To do this seems to require complex modelling of other peoples' beliefs. The next chapter discusses this, and shows that in many cases this need not be as complex as it seems.

Chapter 6. Belief and mutual belief

When people talk to each other, they know that they don't agree on everything. If they did, they wouldn't need to talk, at least, not to exchange information. They know about lots of things in the world. In particular they know about that part of the world which is other people. What they know about is of two sorts. It is facts such as his having dark hair and being married; and it is facts about what the other person believes.

But people know that other people are roughly like themselves. And if they are, then when John talks to Mary, he knows that Mary will have beliefs about what he believes. So he may have beliefs about what she believes he believes. And because he knows she can say the same to herself, we are off on an infinite regress of possibilities for beliefs about what he believes she believes he believes ... How much does this matter in thinking about conversation?

In this chapter I glance at reasons for worrying about belief about others' belief, and then offer a proof of a sufficient condition for mutual belief to arise between observers.

6.1. Why have people thought it mattered?

People do not always understand what the other person has said, but this doesn't always stop conversation. For instance, failure of reference

A: Can you give this to John?

B: Yes, but I won't see him for a couple of days.

A: No, I mean John at work.

Any account of what is going on during this in terms of what A and B are doing must be sensitive to the fact that they do not ascribe the same meaning to B's utterance - or at least, not to start with. Somehow A does come to see into B's mind, or he would be unable to correct him. More important are deception and irony. I attempt to describe them below.

6.1.1. Deception

The whole point of deception is that there is a distinction beteen

- what the deceiver believes, and will found his actions on.

- what the deceiver thinks the victim believes, and will found his actions on.

- what the deceiver believes the victim believes the deceiver believes.

The deceiver attempts to get the victim to believe something that the deceiver does not. If the deceiver does believe it, we would speak, not of deceit, but merely of persuasion. If I tell you that Rice Crispies make you big and strong, and intend you to believe it, only the contrast between what I believe and what I want you to believe distinguishes persuasion and deceit.

It is also vital that there is a distinction between what the deceiver believes and what he thinks his victim thinks he believes. The deciever can't try and fill his victim's mind with falsehoods while letting it appear that he himself doesn't hold them. Suppose I gave you arguments, perhaps very compelling ones, that Rice Crispies

were nutritous while at the same time muttering "Loathsome poisonous muck", which I sincerely believed. I might change your mind. I might even have intended to change your mind. But my action would hardly be deceitful. So any scheme for recognizing deceit would have to be able to distinguish at least these three domains of belief.

6.1.2. Irony

Irony is in some ways like deceit. I say something that I believe to be false. But I do not expect or intend to persuade you of what I say, and I don't expect you to believe that I believe what I say. Suppose you drop the eggs and I say "Such grace, such skill". How do you know that I am speaking ironically? You would, I suggest, say that I didn't believe what I said, probably because I couldn't. No-one could believe that dropping eggs requires skill. Further, you would claim that I couldn't have believed my remark to be persuasive, since no-one can hope to persuade other people of the transparently false.

Any system that hopes to to distinguish irony and deceit will need to know about nested belief. The only other distinction I can see between them is the speaker's intention that his remark being believed. But I don't see how one could attempt to discover the speaker's intention without being able to ask whether some advantage followed for him in the case where he got his hearer to believe something he didn't - if that is he succeeded in deceiving him. But this presupposes nested belief.

A test of whether the speaker expected to be believed on grounds of the general credibility of his remark to someone of his victim's background would not be strong enough, since people may attempt to deceive with futile and incredible lies.

One can distinguish irony and deceit by contrasting the beliefs that one supposes the speaker wants to follow after he asserts "S"

	Irony	Deceit
Sp believes	-S	-S
Sp believes Hr believes	-S	S
Sp believes Hr believes Sp believes	-S	S

But despite this, I think one can manage to explain a lot of what happens in conversation without nested belief. The point is that deceit and irony are wildly unusual in conversation. They are salient; they do occur; it can't be explained without them. But they are only a fraction of what we use speech for. I feel reluctant to use its heavy machinery unless one has to.

6.2. Deriving mutual belief from simpler forms of belief

6.2.1. Related work on mutual belief

Belief in other people's belief (recursive belief) matters to machine understanding of natural language. Two main areas need it.

1) Fixing reference: When I say "the third jar on the left" I give you a way of selecting an object in your world. If your beliefs about the world are different from mine, it may select different things for you and me. You may know we have different beliefs and have allowed for this. I may have allowed for your allowance. To do this I must have recursive belief about you. (eg Clark & Marshall, Donellan, Perrault & Cohen)

2) Indirect speech acts: When I say "Have you got your car?" you have to guess that I ask because I hope that you will guess that I would

only ask (in the current context) if I wanted to know if I could ask for a lift; so you see that I want a lift and so give me one. Such recursive plan reconition need recursive belief. (eg Allen & Perrault, Perrault & Allen, Cohen & Perrault).

The work I've mentioned that handles these using recursive belief is elegant. Why object? The problem is not that the accounts are false, but that they are counter-intuitive. They involve a lot of reasoning about what I believe you believe I believe. When one tells people about recursive belief, they say "That sounds nice, but I'm sure I don't do all that!" If that was all, one could say that they did it all, but inaccessibly. But of course sometimes one does have to stop and think about what the other person knows; during misunderstanding or deceit or irony for instance. Since they sometimes can be conscious of considering recursive belief, perhaps when they think they aren't using it, they really aren't.

I believe we use something simpler most of the time, with full recursive belief available if we need it. That something simpler is mutual belief, a special case of recursive belief. This is the same idea, more formally done, that Clark and Marshall present.

Clark and Marshall approach the problem of mutual knowledge as a result of worrying about how speakers manage to refer. (I consider belief, not knowledge, but it makes no difference here, and Clark & Carlson (1982) continue the argument in terms of belief.) They imagine a speaker Ann trying to decide on a description "t" that she can reasonably believe wiil enable her hearer Bob to identify the referent "R" that she has in mind. They present a series of examples in which it become clear that neither the fact

Ann believes t is R

nor the fact

Ann believes that Bob believes that t is R

nor the fact

Ann believes that Bob believes that Ann believes that t is R

guarantees successful reference. They argue that though inventing counter-examples becomes harder and harder, there is no depth at which it become in principle impossible. So what does allow confidence that reference may succeed?

They argue that what one needs is in fact mutual belief, which they define thus (1981:17)

A and B mutually believe that p <=>

A believes that p &

B believes that p &

A believes that B believes that p &

B believes that A believes that p &

....

and then go on to illustrate how mutual belief of this form can found reference.

This still leaves the problem of how mutual belief springs up. This can arise in several ways, for instance from the participant's knowledge that they are both members of a community in which knowledge of certain facts is universal) but the way that matters

here involves an induction.

First, they observe that when for instance you and I and a table are all together in the same place in the right conditions, then we have mutual belief in the properties of the table. They call the knowledge that we have when we are in such a situation as knowledge of our "co-presence". (This also arises in the form of "linguistic co-presence", but the principle is the same.)

Secondly, they adopt a "mutual belief induction schema" which is

> A and B mutually believe that p iff some state of affairs G holds such that
>
> 1 A and B have reason to believe that G holds.
>
> 2 G indicates to A and B that each has reason to believe that G holds.
>
> 3 G indicates to A and B that p.

Then they point out that in fact co-presence of A and B with the state of affairs p is an instance of G that satisfies the conditions; so mutual belief will arise. This reduces the problem of safe reference to that of ensuring some sort of co-presence with the referent and the hearer.

This seems to me to be wholly right as far as it goes. The ground for mutual beliefs that I offer are very similar to the conditions for their schema. What they do not do is show that the definition of mutual belief that they accept actually follows from their mutual induction schema.

What I propose is to alter both the definition of mutual belief and

the induction schema, in ways that still capture the intuitions we share, but so that the schema really does support the defined version of mutual belief. The important change is in the definition of mutual belief, though it is more one of form than of substance.

Clark and Carlson (1982) go on to discuss the generalization of mutual belief from two people to a group of people, a notion that they call joint belief. Here each person does not merely have mutual beliefs with each member of a group; rather, each person believes that the members of the group collectively know that the members of the group collectively believe some fact. They show that some requests which depend on several people acting in concert can only reasonably be complied with if all the requestees have joint belief. If any requestee doubted that the others shared the joint belief, he would not act, since others might not, and if any of them failed to comply, the joint act would abort. Again, this argument seems compelling. However, neither I nor they can put it into formal shape, nor indeed even define joint belief.

6.3. Deriving mutual belief from simple belief

First some terms. If AbF means "A believes F" then expressions such as AbBb.., AbBbAb..., identify A's belief spaces - the set of sentences to which that expression can truly be prefixed. The hard case of recursive belief is where, because of some twisted story of observation, guesses and error, some spaces contain F and some contain -F. To disentangle deceit or misunderstanding, this is essential.

The easy case is mutual belief. Here either every space contains F or every space contains -F. In the classic example, it is the sort of belief that arises about a lit candle that stands on a table over

which we look at each other.

How does mutual belief help language understanding? I claim that in ordinary, non-ironic, non-deceitful, non-erratic speech, one does not have to work out how what has been said affects different belief spaces. Instead, what follows from the discourse and context can become mutually known. Then to join in or understand later conversation I need consult only one space, the space that contains what is mutually known.

One can't define mutual belief between A and B that F as

mb1(A,B,F) <=>

AbF & AbBbF & AbBbAbF & ...

& BbF & BbAbF & BbAbBbF & ...

because none of us can be sure that we have mutual belief with anybody else. All that one person can do is believe that he and someone else mutually believe something. If one defines it as

mb2(A,B,F) <=> AbF & AbBbF & AbBbAbF & ...

there is a snag. The definition mb2 starts with AbF. But in the candle example, A could say afterwards that though he had relied in what he had said to B on himself and B mutually believing that there was a candle on the table, in fact all along he had believed that, unknown to B, it was really a subtle electrical fake. So I define mutual belief as

mb3(A,B,F) <=> AbBbF & AbBbAbF & ...

This is an infinite expression, and the description of A and B and the candle and their perceptions is presumably finite. How can mutual belief arise from finite grounds? Here is an attempt to show how. I give the proof because though it is ordinary natural deduction form, it uses an intensional logic with non-standard rules of inference.

The essence of the proof is that there are situations which force predictable beliefs on people in them. A sighted man who looks at a candle must believe that there is a candle there. Then, using some brief conditions on someone being in such a situation, and knowing that he is in it, he can perform two inductions about what he and his perceiver think, which together establish mutual belief.

The proof tries to capture the idea of one person reasoning as he thinks another person will. The whole proof is done by one person (here, A): Sometimes A tries to reason as he thinks B will. When he tries to do this, he starts a stretch of embedded proof. A stretch of such an argument is marked by

```
As B
|
| ...
|
End as B
```

Anything that occurs inside that stretch is something that A believes B believes, just as AbBbF states that F is something that A believes B believes. Two rules exploit this. A third rule exploits the assumption that if a person believes F, then he believes he believes F.

Rule "in"

If XbF occurs in some stretch, and that stretch immediately contains another stretch labelled with "As X", then one may write F in the inner stretch, on the same assumptions.

Rule "out"

If F occurs in some stretch labelled with "As X", and that stretch is immediately contained in another stretch, then one may write XbF in the outer stretch, on the same assumptions.

Rule "+b"

If F occurs in a stretch of reasoning labelled "As X", one may infer XbF, with the same dependencies.

If F is a theorem, it may be introduced into any stretch.

Sx means "person x is in situation S". (AbBb)^n means that prefix is repeated n times.

The columns are: the line number; a sentence; the rule and lines used to derive it; the assumptions on which the line depends.

PROOF

As A. The assumptions:

1	y(Sy -> ybF)	A	1
2	Bb(y(Sy -> ybF))	A	2
3	SB	A	3
4	BbSA	A	4
5	BbSB	A	5

I abbreviate the conjunction of these assumptions as <gmb>, "grounds for mutual belief".

6	<gmb>	+&, 1,2,3,4,5	1,2,3,4,5

Base case 1

As B

7	SA	in, 4	4
8	y(Sy -> ybF)	in, 2	2
9	SA -> AbF	- UQ, 8	2
10	AbF	- ->, 7,9	2,4

End as B

11	BbAbF	out, 10	2,4
12	AbBbAbF	+b, 11	2,4
13	(AbBb)^1 AbF	rewrite 12	2,4
14	<gmb> ->(AbBb)^1 AbF	+ ->, 6,13	-

So <gmb> ->(AbBb)^1 AbF is a theorem.

Base case 2

15	SB -> BbF	-UQ, 1	1
16	BbF	- -> 3,15	1,3
17	AbBbF	+b, 16	1,3
18	(AbBb)^1 F	rewrite 17	1,3
19	<gmb> -> (AbBb)^1 F	+ ->, 6,18	-

So <gmb> -> (AbBb)^1 F is a theorem.

Induction 1. The induction hypothesis is that

<gmb> -> (AbBb)^n AbF is a theorem.

```
As B
|
| 20    y(Sy -> ybF)              in, 2          2
| 21    Bb(y(Sy -> ybF))          +b, 20         2
| 22    SA                        in, 4          4
| 23    SB                        in, 5          5
| 24    BbSA                      +b, 22         4
| 25    BbSB                      +b, 23         5
| 26    <gmb>                     +&, 20,21,23;24,25
|                                                2,4,5
| 27    <gmb> -> (AbBb)^n AbF     Theorem        - (I.H.)
| 28    (AbBb)^n AbF              - -> 26,27     2,4,5
|
End as B

29      Bb(AbBb)^n AbF            out, 28        2,4,5
30      AbBb(AbBb)^n AbF          +b, 29         2,4,5
31      (AbBb)^n+1 AbF            rewrite 30     2,4,5
32      <gmb> -> (AbBb)^n+1 AbF
                                  + -> 6;31      -
```

Which proves the induction step. So, with base case 1,

<gmb> -> (AbBb)^n AbF is a theorem for n >= 1

Induction 2. The induction hypothesis is that

<gmb> -> (AbBb)^n F is a theorem.

The proof is exactly the same as for induction 1, except that "(AbBb)^n F" replaces "(AbBb)^n AbF". So, with base case 2,

<gmb> -> (AbBb)^n F

is a theorem for n >= 1

So <gmb> -> (AbBb)^n F & (AbBb)^n AbF

is a theorem for n >= 1.

That rewrites as AbBbF & AbBbAbF & ..., which is the definition of mb3(A,B,F). So if A believes <gmb> for some S,F, then A believes he and B mutually believe that F.

The stronger case, mb2(A,B,F), can be derived by adding the assumption "SA" which will not figure in the induction but which will entail the first conjunct in the definition (AbF) as well as all those that also occur in mb3.

Going back to the application to natural language: "Sx" will mean "x was in places where x must have observed certain context and heard certain parts of the discourse; and x has a theory Theory for analyzing discourse in context." F will then be the deductions made using that theory. What I have said about mutual belief places no constraints on what Theory is. It just says that for the parts of the discourse, context and theory that fit the conditions for mutual belief, the conclusions will be mutually known. That means that one may be able to fillet the parts of your favourite theory of the bits of it that handle recursive belief, leaving something perhaps simpler behind.

6.4. In defence of mb3

The definition of mutual belief (mb3) that I am using is contentious. It is altered from the customary definition (mb1) in two ways - by dropping all the conjuncts starting "Bb..." (mb2) and by dropping initial conjunct AbF (mb3). The latter change is less dubious; I have defended it earlier; but even if those arguments are rejected, only trivial alterations are needed to the proof of how mb3 arises to convert it into a proof of how mb2 arises. One need just add "Abf" to the grounds for mutual belief. Given that, "why is mb3 better than mb1?" amounts to " why is mb2 better than mb1?", since any answer to the second can be converted into an answer to the first.

Mb2 has its supporters: Joshi (1982:183) uses it, where he calls it "one-sided mutual belief" (and mistakenly identifies it with the definition of Clark & Marshall (1981)). But he does not explicitly defend it. There are however two arguments for it. The first is that it is forced on us unless we are to make excessively strong claims about our knowledge of other minds; the second, that it has to be of that form to be part of any teleological explanation of action.

6.4.1. Argument 1

Suppose one is trying to construct an induction schema that can explain how mutual belief can arise. Then what one is doing is trying to prove a theorem of the form

<grounds> -> mb(A,B,F)

The grounds may mention both the world and A's and B's beliefs about the world or other beliefs. When proving this theorem, one can take a god's-eye view and postulate what one likes about exactly what it is that A and B know. But this theorem is not going to be used by a god; it is going to be used by an agent, A say, to deduce that if he and

someone else are in such-and-such a situation, and so satisfy <grounds>; then they have mb(A,B,F). But this lets in a sceptical argument.

What does A's belief that the grounds are satisfied entitle him to assent to? Not "<grounds> are true", but "I believe <grounds> are true". In general one might rely on some standard counter to a sceptical position to say that these are equivalent. Such a counter may be usually true but it can't be used here, because <grounds> refer among other things to B's beliefs. If <grounds> include a conjunct "B believes F", then if the anti-sceptical counter is to work, it entails the equivalence of "B believes F" and "I believe B believes F". That cannot be accepted by anyone unless he holds

- both that, in principle, other minds are knowable as certainly as the physical world is knowable

- and that, in general, (whenever the induction schema is to be used, at least), agents know the contents of each other's minds as well as they know the physical world.

Arguments for those claims exist - they are given by at least behaviourists, people looking forward to a complete reductive neuropsychology, and some ordinary-language philosophers concerned with the use of "believe". However I shall assume their falsity here.

If though one does take "F" said by A to be equivalent to "A believes that F" the schema can still be used. A can still reason

```
A believes <grounds>
A believes (<grounds> -> mb(A,B,F))
-----------------------------------
A believes mb(A,B,F)
```

but the version of mutual belief he will emerge with is that that I am seeking to defend as the correct analysis.

The reason for that is that I think the definition of mb(A,B,F) will have to be either of the customary, mb1, form

AbF & AbBbF & AbBbAbF & ...
& BbF & BbAbF & BbAbBbF & ...

or of the mb2 form

AbF & AbBbF & AbBbAbF & ...

In either case, what A can conclude is Ab(mb(A,B,F)). For the mb1 definition,

Ab(mb(A,B,F))

<=> A b (AbF & AbBbF & AbBbAbF & ...
& BbF & BbAbF & BbAbBbF & ...)

distributing Ab... over &

<=> AbAbF & AbAbBbF & AbAbBbAbF & ...
& AbBbF & AbBbAbF & AbBbAbBbF & ...

reducing iterations of AbAb... and removing repeated conjuncts

<=> AbF & AbBbF & AbBbAbF & AbBbAbBbF & ...

In the mb2 case, by the same steps,

Ab(mb(A,B,F))
<=> Ab (AbF & AbBbF & AbBbAbF & ...)
<=> AbAbF & AbAbBbF & AbAbBbAbF & ...
<=> AbF & AbBbF & AbBbAbF & ...

<=> mb(A,B,F)

That is, whichever definition of mutual belief is used in the theorem that is the essence of the induction, what A can conclude is of mb2 form.

6.4.2. Argument 2

Suppose one undertook to explain an agent's rational actions in terms of his beliefs and values. Then obviously one is only allowed to refer to what the agent believes to be true, not to what is true. If Fred is known to want to get to town, and is seen to walk to a bus stop, one can say "Fred is going to the bus stop because a bus is coming" only as an abbreviation for "Fred is going to the bus stop to catch a bus because HE BELIEVES a bus is coming".

Mutual belief is to be used as a notion that helps explain agents' rational actions. But if it is used in this way, then, just as with beliefs about whether a bus is coming, what matters is not that two agents have mutual belief, but that the agent whose actions are to be explained believes that they have mutual belief.

Teleological explanations of A's actions (or B's - it doesn't matter) depend on the existence of arguments of the form

```
A wants F
A believes G
A believes act Act done when G entails F
-----------------------------------------
A does Act
```

where, in the cases in which mutual belief about some fact are explanatory, G will be of the form

P & mb(A,B,F)

where P is everything else that it matters that A believes. So if one is to give explanations of the form above, one must show

Ab (P & mb(A,B,F))

<=> AbP & Ab(mb(A,B,F))

Concentrating on the conjunct that matters to this argument, one has to show

Ab(mb(A,B,F))

But by the steps given above, that is equivalent in either case to mb2(A,B,F). So for a teleological explanation to go through, what one has to demonstrate is always mb2(A,B,F).

6.5. Application to conversation

How does this apply to utterances seen as attempts to affect plans? The central notion is that one can sometimes forget that the other's plans and beliefs and values are concealed in his head, and that yours are concealed in your own, so that both of you have to make fallible models of what is inside the other. Rather one can in many cases act as if all your important plans and beliefs are as it were laid on a table between you, so that you both know what you each think. That is, the plans etc are mutually known. Furthermore, what each of you says may be able to change what is on the table. If the conditions arise for the change that you make to be mutually known, then what is mutually known will be changed. I claim that what we do to our plans in conversation is often like this, and that we often have no need to consider nested belief.

How does this happen? The idea is that at some point in the

discourse, certain of the parties' plans are already mutually known. Then one of them makes a remark. That remark will change the plans. Into what? Well, assuming the remark was rational, it must secure some good. Suppose one can find an account of how the plans must be changed if that good is to arise. The speaker must suppose that he has made that change, and that some of his hearer's beliefs etc must be different.

Recognizing the benefit of the change a remark makes to a set of plans may involve ascribing new or altered plans to some of the speakers. But the methods of recognizing the benefit are standard. Everyone who talks can, I suggest, use them. Anyone who knew the initial plans etc and the remark made will know the resulting plans. But the plans they were applied to were mutually known to start with. And the remark will be mutually known - all the parties can see that the others heard it, and know that they must admit they heard it. So what everyone believes after the benefit of the remark has been found will be mutually known too. Consider the exchange:

I: I want an apple. Shall I get you one?

You: The kitchen floor is wet.

I've told you I want to eat an apple. Certainly I may have a plan about how I am going to do this, perhaps involving going to the fruit bowl in the kitchen. But if I do, you know nothing about it.

Nevertheless, I am able to say things that you both ought to, and will, connect to that plan. For instance "Shall I get you one?" You will understand that my offer is bound up with my own plan, and you will do this although you have no detailed account of what I am going to do other than that it ends in my having an apple. What happened was that we started with a mutually known plan, that I should get

myself an apple. This was mutually known because I had declared my goal, and what we state is mutually known. This goal was a degenerate plan; degenerate in that nothing was known of it except its aim. The means were unspecified. Then I say something that can be "hooked into" what I have already said only on the assumption that I intend to go and get myself an apple; rather than say asking you to go and get me one. But if this assumption is made, the plan is more fully specified, and if we both know about it in its fuller form, we can both talk about it like that. For instance, you say that the kitchen floor was wet. This carries the suggestion that I shouldn't walk on it. This only has significance because we both know that the plan we have had to ascribe to me involves me walking on the kitchen floor. We both know the plan that we are talking about, even though it has never overtly been described.

My intuition about what happens in the exchange about wanting an appple is this. At each stage, we say "This has been said. By canons of explanation of action that I think we share, I think this suggests that the plan he has in mind is so-and-so. But since I made this deduction, and all I used were publicly available facts and common canons of explanation, if I came to this conclusion then I can think that he did too. So the plan I have ascribed to him is known to us both as a rational interpretation of the foregoing discourse".

Imagine an anthropologist reading the text of a conversation between two people. Assume he knows about the participants' material and social culture. He may well be able to understand the purposes of each utterance, just as you understood the puposes of the utterances in the apple dialogue above. At any moment he would have been able to say what plans he had had to ascribe to the speakers on the ground of what they had so far said. I suggest that this approach is open to the speakers themselves. Each of them will think of himself and his

interlocutor as a skillful anthropologist, so each of them will suppose that the other has come to the same conclusion as himself. So each can talk about anything he has deduced, with the expectation that the other will understand him.

6.5.1. A slightly more formal account of how plan changes are mutually known

Slightly more formally: suppose the parties to a discourse are A,B,... Then what one of them takes to be mutually believed (perhaps about the world, perhaps about others), intended and wanted before the remark U is

A believes ... & A wants ... & A intends ...
& B believes ... & B wants ... & B intends ...

Call this "X before U". Suppose that we also have an axiomatized theory that tell us how peoples' beliefs will change after a remark, given that they will only accept such a change if they can see how it benefits the remark's maker. This theory would be the real version of the theory I am trying to outline in this thesis. Call it Theory. Then one could use it to infer "X after U", which will be another description of the parties' plans etc after they have found an explanation of why the speaker spoke.

X-before-U & Sp said U & Theory -> X-after-U

If one accepts that from A -> B one may infer y(ybA -> ybB) then I may say

y(yb(X-before-U) & yb(Sp said U) & yb(Theory) -> yb(X-after-U))

But this matches the vital step in establishing mutual knowledge, with

Sx <=> yb(X-before-U) & yb(Sp said U) & yb(Theory) F = (X-after-U)

Do the other conditions also hold? S is a three part conjunction. Do the conditions hold of each part?

X-before-U: we are talking only about the mutually known beliefs etc of the parties. From the definition of "mutually known", the other conditions must hold of X-before-U.

Sp said U: Whether the conditions on this are true will be a matter of fact that will vary from occasion to occasion. Certainly it will follow if it is mutually known that the remark was made, which is the usual case.

Theory: If A believes that we all apply Theory to explain a speaker's utterances, and if he believes that is is mutually known that we do this, then the other conditions follow. This assumption can be defeated. We all have met people who respond to indirect requests differently to how we do, and we may not assume that we and they mutually know Theory. But I claim that in general we may.

If this goes argument goes through, I claim it permits me to treat all knowledge that participants in conversation have as mutual, as long as nothing goes wrong. The actual utterances that the speakers made are mutually known in the way that their knowledge of their surroundings is mutually known (though strange things can happen if they ever disagree about what they said). If the means of recovering intention from utterance are mutually known, then the ascribed intentions are mutually known.

What does the "as long as nothing goes wrong" qualification mean? In some cases, an utterance can be connected to an intention in more than one way. Then one of two things happens:

The participant must explicitly keep in mind that it is mutually known that one of two plans is being mooted or executed, and each subsequent remark must be considered in two lights. Perhaps a later remark will kill off one plan if it turns out the it can connect to only one plan. For instance

A is about to go shopping

B: Do you have enough cash?

A: (wonders: Is he going to offer to lend me some? Or ...)

B: Do you have enought to let me have some too?

A: (thinks: Yes! he wants to borrow some)

6.6. Joint plans

Who owns plans? When we talk to each other, the immediate picture one has it that you have your plan or plans, and I have mine. We may tell each other about them, we may assist each other with them, but each plan will belong to exactly one of us. I am not sure that this is always the best way to look at them. Consider this example:

A: I want the Sunday newspapers.

B: I'll get them.

A: I'll give you some money.

The plan that should be formed as the result of this exchange involves actions by two agents, so it can't belong just to its agent.

Nor is its goal something that will please just A or just B. Either B wants A to be able to read the papers, and therefore adopts getting the papers as a goal of his own; or perhaps B already wanted to read the paper, and is going to buy the newspaper knowing that both he and A will benefit. So it may be hard to say who a plan belongs to just on the grounds of who wants what it achieves.

I believe that a better approach is to suppose that a plan can stand in front of several speakers so that they can operate on it in sight of each other just as they could on operate on a car engine standing before them in a garage.

This is probably the right way to view all plans. They do not belong to any one in particular. Rather several plans are mutually known to exist, and what is known about them is not just their structure, but also who is expected to perform them or benefit from them. Otherwise, the plan that lies behind A saying

A: Do you want an apple?

or B saying

B; Do you want an apple?

are going to pose excessive nice questions of who the plan belongs to, performed as they will be for another's good.

Certainly one has to see a plan as joint when one person has revealed or discovered a need derived from it, which the other both can and will assist with.

Chapter 7. Series of actions on plans

Certain series of alterations to mutually known plans count as the performance of speech acts; it is useful to speakers to be able to see such series.

- if they know what another has just been doing, it may suggest what he is doing now. I want to use the offer-request/accept-reject/refuse-comply nexus as an example of this

- once one has discovered that what has been said has a useful relation to what has gone before, one needn't go on looking for the reason it was said. I want to use the giving of explanations as an example of this

7.1. Relations between "speech acts"

Speech acts are actions that we perform by speaking - eg warning, threatening, promising, refusing, christening, denouncing. I want to claim two things about them.

1 - To define some of them, one has to talk about the speakers' and hearers' plans and how the utterances that count as that speech act affect those plans.

2 - Speech acts are not just labels. One can explain the performance of speech acts by saying their speaker is performing a plan of which they are part. Such a plan will have the same form as a plan to act in physical world, but will act on different sorts of thing and involve different actions. The objects acted on in such a plan will be other plans. If one refuses to admit that there are such plans,

then there will be speech acts whose definition one can't give and whose use one can't explain.

Furthermore, there are words that describe plans. We use them to tell and ask each other about their plans and ours: I mean words like "can" and "shall". Just as (some) speech acts are actions on plans, these words are (sometimes) predicates over plans. This matters because by using them we can alter what is known about our plans, perhaps with great effect.

Let me give two examples.

1 - I am just coming up to the door of the department where I work. A janitor has just left, closing the door after him. He sees me and says "Do you have your keys?". He is offering to open the door for me if I do not have them.

2 - I come up to my front door where I see my wife is already standing. As I get there, she says "Do you have your key?". She is requesting me to open the door for both of us.

The utterance is the same. The speech act performed is different. The difference must lie in the context. Where?

One account of what a request is, is that it allows Hr to infer a want of Sp that Sp cannot himself achieve but which Hr can. That definition has to be tidied up. I can express a wish eg to have a hundred pounds, that you could fulfill and that I really want, but which I do not expect you to fulfill for me. (I might even be embarassed if you tried to.) Sp's request has to be about a want that Sp believes that Hr will get for him once Hr knows that Sp wants it. I have no criteria for that, though clearly they include the

detriment to Hr, the benfit to Sp, and the goodwill of Sp towards Hr. (There is also a difficult social quantity: one person may want to accept a disproportionate request so as to build up credit for his own later request.)

If a request is the revelation of a want, what is an offer? One could say it was the revelation by Sp of his ability to do something that Hr may want. But again, Sp could reveal that he could lend Hr a hundred pounds without that counting as an offer. And where is the symmetry between offers and requests?

The distinctive and unifying features of requests and offers is that each make it mutually known that one party contemplates a plan that involves actions by one of them to help the other. In a request, Sp makes it mutually known that he contemplates a plan that benefits him, which inculdes actions by Hr. In the door example, my wife makes it mutually known that she contemplates getting into the house by using a plan in which it is me who opens the door. In an offer, Sp makes it mutually known that he contemplates a plan that benefits Sp which includes action by himself. In the door example, the janitor makes it mutually known that he contemplates my getting into the department by a plan in which it is he who opens the door.

A and B have gone shopping together.

A: I've left my purse behind.

B: I've got my cheque book.

There is a request-and-offer feel to this exchange. Why? First, A impugns something that might be his own plan - he won't be able to pay for the shopping. But if he can expect that B will help him with what he can't do himself, and if he thinks that B is able to help

him, then what he says is a request. He contemplates a plan in which B's actions help him.

A doesn't have to have a specific expectation about how B will help him. He doesn't (say) have to hope for B to use a cheque rather than a credit card; just so long as it something B is to do.

A has to make what he contemplates mutually known. His attack on his own plan succeeds in this. How? Because A and B mutually know that another way A could obtain his end is by explicitly asking B to help him. They both know that when a plan collapses, one tries to mend it. So they mutually know that A must be thinking of mending it. If they both know that getting B to help is the best way of mending it, they mutually know what A must intend, and can react to that without further discussion.

But if they are so secure in what they know, why does B go on and say that he has his cheque-book? Because they both know that they may not in fact have duplicated each other's reasoning.

7.2. Sequences of speech acts

After a request or an offer, Hr will respond.

After a request, Hr can COMPLY:

A: Can you lend me $5?

B: I'll just get my wallet.

A: Will you give this to Celia?

B: Where will I find her?

Or he can REFUSE:

A: Can you lend me $5?

B: I'm afraid I'm broke.

A: Will you give this to Celia?

B: I won't be seeing her.

After an offer, Hr can ACCEPT:

A: Would you like some tea?

B: I'd love some.

Or he can REJECT:

A: Would you like some tea?

B: I just had some, thanks.

How can we tell which, if any, of these a remark is? I claim that one has to look at two sorts of facts

- what changes the remark makes to mutually known plans

- when this change occurs relative to other changes

That is, something is a compliance (say) only if

- it does certain sorts of things to the parties' plans.

- the parties plans are as they are because of a request.

In the compliance cases; B makes it mutually known that he is getting

his wallet. This action is explicable as part of his execution of a plan to lend A $5. So he must intend executing the plan A requested. Similarly with the second example. The information B asks for is only useful if he intends giving whatever-it-is to Celia as A requested.

In the refusal case, what B does is show why he won't perform the plan that A requested. In the examples, he says things that would make the plan impossible.

In the acceptance case, B encourages A's contemplated plan by showing reasons why he should perform it: that B wants it.

In the rejection case, B discourages A's plan by showing that B doesn't want it.

What is the generalization? I take it to be this:

A request by A to B to do X is an action (usually an utterance) that makes it mutually known that A contemplates a plan
- to achieve one of A's own goals, G,
- part of which is B's doing X,
- when previously B's doing X was not known to be part of the plan.

A compliance and a refusal presuppose an earlier request.

B complies with A's request to do X if B does something that makes it mutually known that B contemplates doing X to achieve G. Similarly, B refuses A's request to do X if B does something that makes it mutually known that B does not contemplate doing X to achieve G.

Offers, acceptances and rejections are symmetrical with requests, compliances and refusals.

An offer by A to B to do X is an action (usually an utterance) that makes it mutually known that A contemplates a plan

- to achieve one of B's own goals, G,
- part of which is A's doing X,
- when previously A's doing X was not known to be part of the plan.

Again, an acceptance and a rejection presuppose an earlier offer.

B accepts A's offer to do X if B does something that makes it mutually known that B contemplates achieving G by a plan that includes A's doing X. Similarly, B rejects A's offer to do X if B does something that makes it mutually known that B does not contemplate achieving G by a plan that includes A's doing X.

The example above fit these definitions. But there are some slightly odd cases:

When a particular action is requested, the requestee may refuse that action but offer another.

A: Will you give me a ring when you get home?

B: I'll ring you before I leave work.

A's goal G is that B get in touch. The requested action X is that B ring when he gets home. But B declares that he will do something else, Y, to wit, ring before he leaves. If as seems reasonable one can deduce that doing Y involves not doing X, then B's remark is properly both a rejection and an offer.

When A and B have a joint plan, they will have shared goals. But the definitions survive. Suppose A and B are trying to start a car they

hope to travel in. Then

A: I'll borrow some jump leads

is an offer, and

A: Will you borrow some jump leads?

is a request.

One problem is conventional refusals and compliances. Eg

A: I've just spent my last farthing.

B: Tough.

A: Can you give me a light?

B: Sure.

Clearly these are a refusal and a compliance. But this doesn't come by deduction from an explicit knowledge of what B is going to do. One has to know in advance what "Tough" and "Sure" are used as in these contexts. I don't know how such conventions work.

7.3. Utterances that are explanations

Consider some examples:

example 2/67

M and W have a guest, Aileen, staying in their flat. Aileen intends to go out and do some shopping by herself

M: I gave Aileen a key.

W: Why? Aha! Because of her shopping.

example 2/78

M has just turned on the immersion heater

M: I'm leaving the hot water on for my bath.

example 2/93

M and W are just about to start eating dinner

W: I don't like to eat my potato with a teaspoon.

M: They're for the avocado.

All of these examples feel to me as if they have something in common: in each of them one of the speakers is explaining something that has been done. He says what he says so that with the new information the other can fit what is being explained into a plan that he wouldn't otherwise have known about. Once he can fit it in, he has an explanation of it.

I am suggesting that one can explain some utterances as explanatory. One can say "Aha, he said what he did because what he said explains something which it was to his advantage to explain". This leaves me with the tasks of answering the questions

- How do utterances build explanations?

- Why are explanations worth undertaking?

7.3.1. How do utterances build explanations?

I don't want to get embroiled in a lot of pure philosphy, so I'll be dogmatic about what explanations are.

Facts are the only sort of thing that is explained. Objects for instance aren't. I claim that you have an explanation of a fact when you can see that it follows from what you already accept.

This is in general too strong a claim. For instance, you could say that you had an explanation of a particular revolution if you could show that in 80% of cases when a nation was in the state that preceded this revolution, a revolution had followed. But if you then have an explanation, you have it although the explanandum didn't necessarily follow from the explanans.

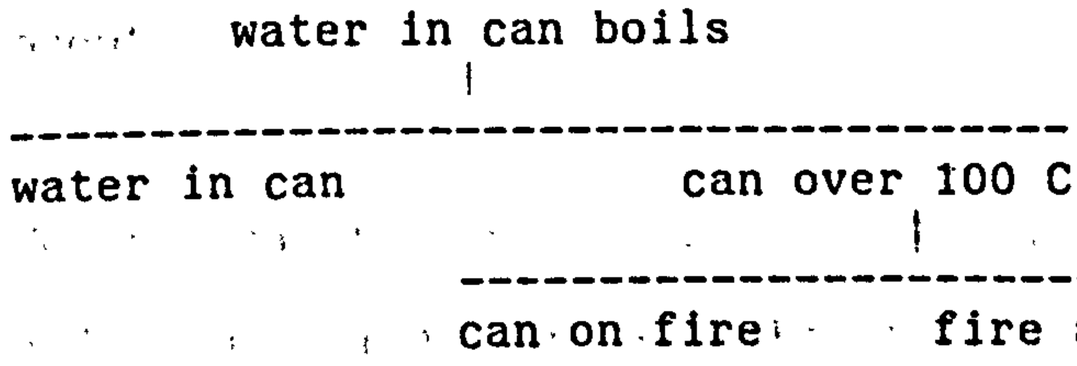
The claim may also be too weak. A fact might follow by some tortuous chain from axioms that you accept, but you may nevertheless object that the length and complexity of the chain prevented it from counting as an explanation. Yesterday's weather and gas dynamics together entail today's rain, but one would hardly say that the calculation that showed this was an explanation of the rain.

So I am wrong about what an explanation is. But on any account, being able to show that the explanandum does follow reasonably simply from what you already accept should count as having an explanation.

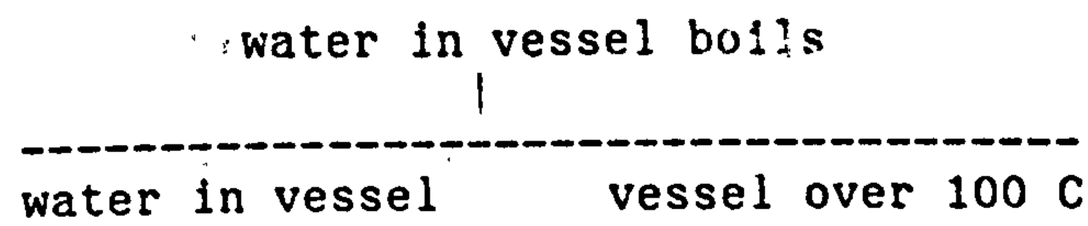
7.3.1.1. Teleological explanations

Given the version of explanation I offer above, the explanation of any physical fact is going to be some deduction of that fact from given facts via known physical laws. For instance, suppose I want to

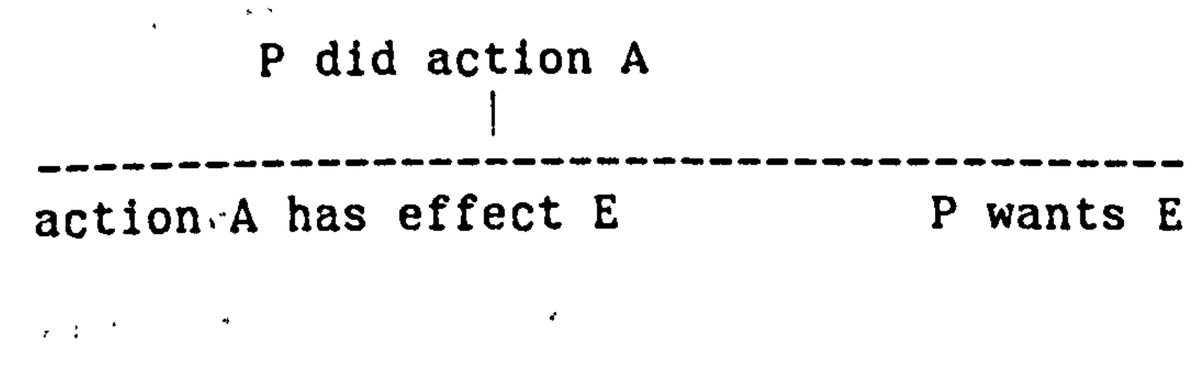
explain why a can of water is boiling. To do that, I offer a proof which can be sketched as in the diagram below. I rely on your intuitions to interpret it. It is an and-tree with the consequence at the top.



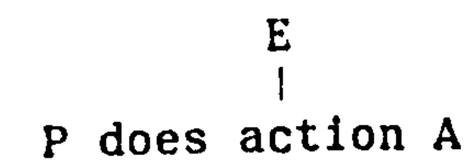
Each and-step is justified by its being an instance of a law which is not written out explicitly. For instance the top step of that proof could be an instance of the law



But if I want to explain something, not according to physical laws, but teleologically, what must I do? To explain something teleologically is to explain it as the result of agents' beliefs and desires. To explain it in the same way as physical events are explained, one would need laws that axiomatize planned behaviour, and facts that describe the agent's beliefs and desires. For instance, an explanation of why a person P did some action A might look like this:



On the other hand, one could explain the same action by exhibiting the plan of which it was a part, and asserting that that plan was being executed by P. The plan would look something like



The form of this argument is different to that I just described for explaining physical events. There a fact was true if it occurred at the top of a proof tree. Now I want to say that certain things are true about a fact (or an action) if they occur at certain sites in a proof tree.

You can say that any node in a plan is true if the nodes below it are all true. You can say that a node in a plan is wanted if the node above it is wanted. In a complete picture of a plan, the top node will be wanted, so all the nodes in the plan will be wanted. If you believe that the plan has been executed, then all the nodes at the bottom of the plan will be true, and so all the nodes in the plan will be true. If you believe that the plan is being executed, then you will believe that some of the nodes will be true already, and some will be made true later. If you know that the plan is going to be executed, then you know that all the nodes may be made true later. For those nodes that are actions rather than facts, "being true" amounts to "having been executed".

So to explain the truth or wantedness of a fact, or the performance or wantedness of an action, is to proffer a plan that involves that fact or value in the right place. Whether the truth is current or anticipated depends on the execution status of the plan.

To give an explanation of an action (say) is to give a plan which explains that action. But to give a plan, it isn't necessary to exhibit it in its entirety. One merely has to make it accessible. Just making the goal known may be enough, if Hr can guess what plan the agent will make to get that goal. For instance,

A: I'm going to Bauermeister's <a bookseller>. I want a Spanish dictionary.

7.3.1.2. Explaining omission and explaining commission

All that I've just said applies to explaining why something is the case or was done. But equally one may want to explain why something wasn't done or wanted. For instance,

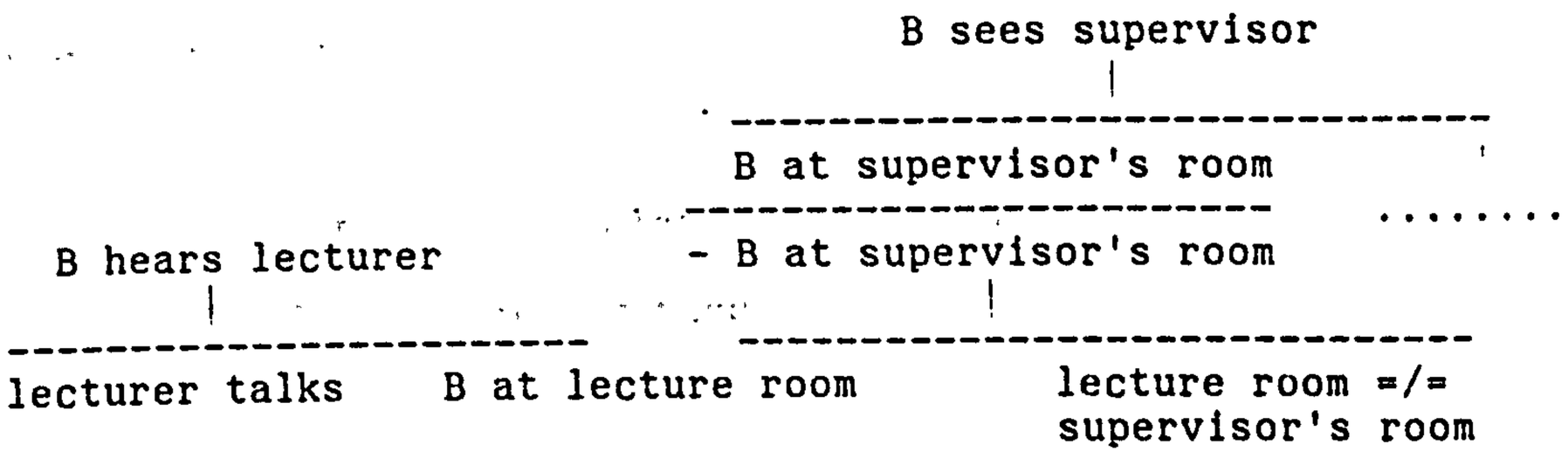
A: Why didn't you come to the seminar yesterday?

B: I had to go to see my supervisor.

Just as someone explains the occurrence of some event to you by showing that there is a plan that explains it which you didn't know about, he can explain its absence by showing that the plan that contained it and which you thought was going to be executed in fact won't be. In the example above, A thought B had a plan roughly like this:

```
          B hears lecturer
                 |
-----------------------------------
lecturer talks   B at lecture room
```

But B says that he has a different plan. This other plan was such that the preconditions it needed gave the lecture-going plan bad side-effects. And because the lecture-going plan is now known to have bad side-effects, A can no longer be surprised that B didn't pursue it. The omission of the action is explained by the abandonment of the plan.



Giving explanations is a matter of affecting the picture that someone has of someone else's plans. The diagram below summarizes the relation between the type of explanation and the change made.

	known plan of agent, as it is:	
	before remark	after remark
Explaining comission	contains explanandum	doesn't contain explanandum
Explaining omission	doesn't contain explanandum	contains explanandum

7.3.1.3. 'Explaining how' versus 'explaining why'

One doesn't need to construct a brand new plan around the thing to be explained in order to give an explanation. Anything that is capable of changing our picture of people's plans may do. All that has to happen is that the explanandum, not formerly known to be part of a plan, ends up as part of a plan.

This shows up in the contrast I'm going to make between explanations of how something is done with explanations of why it was done at all. Consider the examples:

example Spanner-1

A: I want a spanner. I'm clearing the U-bend.

example Spanner-2

A is known to be clearing a U-bend, and has been seen trying to get the nut off by hand

A: I want a spanner. The nut is stuck.

In Spanner-1, A gives an explanation for his want by revealing a plan that he wasn't previously known to have. His want forms part of it. In Spanner-2, his utterance kills off one plan (to undo the nut with his fingers) by showing that it won't work. One can only undo nuts that aren't stuck that way. He will do it instead with a spanner. He doesn't say this, but if you look round to see how he will do it, you will light on the plan of him doing it with a spanner. Hence his want.

In cases like these one is offering reasons for choosing between different means to one end. Any sort of reason can be offered. One can give reasons for dropping one alternative. It may have bad side-effects:

A: I'm going by car. It's pouring with rain.

Or be unsound:

A: I'll make white bread. I'm out of wholemeal flour.

One can equally give reasons for adopting the alternative because of its own merits. It could have good side-effects:

A: I bought these at Fraser's. There was a sale.

Or because it is sound. In this case these seems to be a suggestion that the altenatives wern't sound. Perhaps to defend a choice of means one has to show that no other alternative was as good. If you think that someone is defending his choice of means in this way, but you can't see him explicitly disparaging the alternative, you will presumably accept that he thinks he could disparage it if he wanted. So in

A: I bought this at the Indian shop.

where A's choice lay between buying whatever it was at the Indian shop, or somewhere else, you will expect him to be able say what was wrong with buying it somewhere else. Perhaps everywhere else was shut, in which case the alternative to what he did would have been unsound.

7.3.2. What sort of things are explained?

If one is going to try and explain an utterance as explanatory, one needs to say what it is explaining. Choosing the candidates is a bit more complex than it may initally seem.

7.3.2.1. Explaining statements

Many of the things that one wants to explain teleologically are actions. One sort of action is what one could call action out-in-the-world: picking up and putting down, coming and going. To explain this one reveals a a plan of which the action is part, and which will presumably be intended to achieve some out-in-the-world benefit - having something or doing something.

Utterances are actions, and are themselves candidates for

explanation. But look at these examples:

example SX-1

A: There's not much money in your account. I don't want you to become overdrawn

example SX-2

A: There's not much left in your account. This month's pay cheque hasn't been paid in

example SX-3

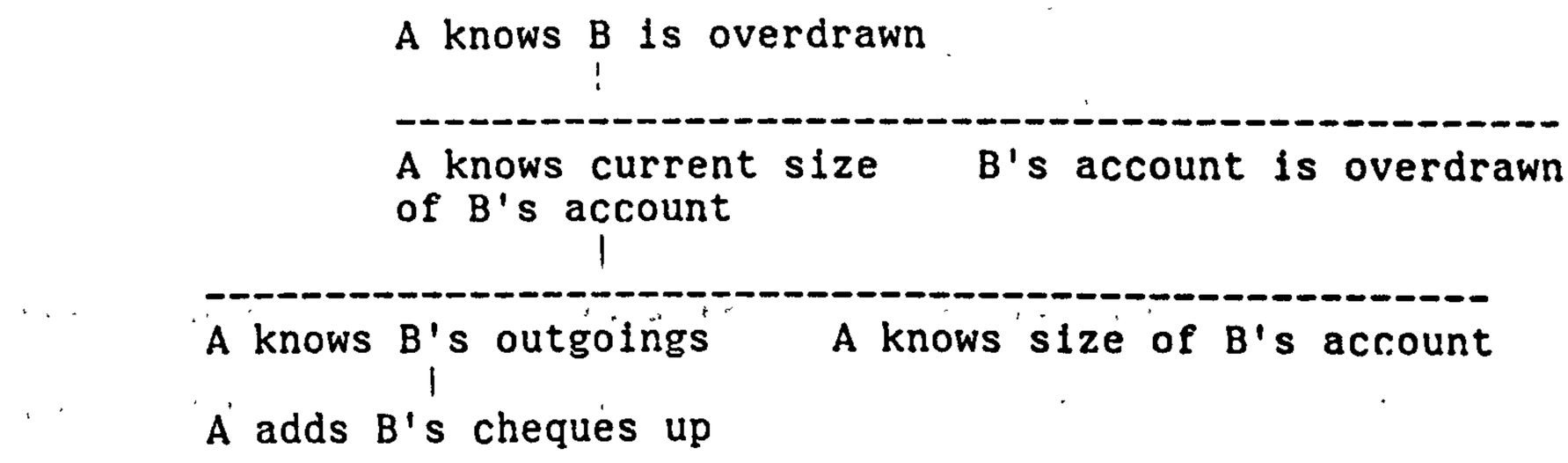
A: There's not much left in your account. I've just added up what you've paid out last month.

In all of these, the second sentence is offering an explanation germane to the first sentence, but in different ways.

- In SX-1 one interpretation of what A says is that he is explaining the reasons for his saying the first sentence by means of the fact asserted in the second. A's first sentence, if construed as a warning, should be explained as part of his plan to get B to abandon a plan that would lead to B ending up overdrawn. A does this by showing that he does have such a plan. He must have it because of the goal that he announces he has: "I don't want you to become overdrawn".

- In SX-2 A is explaining is the fact asserted in his first sentence, of someone not having much money in his account, by another fact which from which it follows.

- In SX-3 A is explaining, not the fact, but the fact that he knows the fact. He has mentioned an action which is involveds in a plan something like this.



Similarly, one can explain one's ignorance of a fact. Asking a question is usually grounds for thinking that the asker doesn't know the answer, and explanation may be offered for that fact.

On a railway station platform

A: Did you get the announcement? I couldn't hear it for the noise.

7.3.2.2. Explaining requests

Questions are the companion of statements. They should be explicable too. I'll start by talking about explaining them, and then suggest that it is requests rather than questions that we should see as the companion. Questions are after all a particular class of requests.

Questions show some of the same ambiguity about what exactly is to be explained that statements do, but it is not quite the same. If they were the same, the parallel would look like this.

	statement	question
thing to be explained	why did he say F?	why did he ask F?
	why is F true?	why is F wanted?
	why does he know F?	why does he not know F?

But there are problems with this. If one asks "F?", it is not true that one wants F. What one wants is to know whether F. So the second line is funny. Then again, why does one ask a question? Because one has a reason for knowing something, but one doesn't know it. But then the second of these is both a component of the reason in the first line, and also the whole of the third line. If instead I draw a parallel between statement and request, the table looks like this.

	statement	request
thing to be explained	why did he say F?	why did he request F?
	why is F true?	why is F wanted?
	why does he know F?	

Are there two distinct cases? Yes. Contrast the two examples:

A: Could you give some people a lift over here? I ask you because you've got the largest car.

A: Could you give some people a lift over here? We need them to complete our octet.

Only in the second is a reason given for wanting the people to come over at all.

It is still true that asking a question can lead to an explanation being given for the asker not knowing the answer. But that this can

happen can be extracted as a theorem in this setup. A question is a request to tell something, say G. Being told G is only profitable if you want to know G, and you don't know G. So to most people who ask a question, you can ascribe a desire to know G. And then that desire is available as a thing to be explained.

In fact, something similar may be true for all requests. One only requests X if one is actively seeking X. One only actively seeks X if it both something you want and if (one knows) that it is false. So any requests suggests that both -X, and knowledge of -X, are available for explanation.

If you view a statement as a revelation of what Sp believes, and a request as a revelation of what he actively seeks, then the proper analogy will be this.

	statement	request
thing to be explained	why did he say F?	why did he request F?
	why is F true?	why is F valued? why is -F true?
	why does he know F?	why does he know -F?

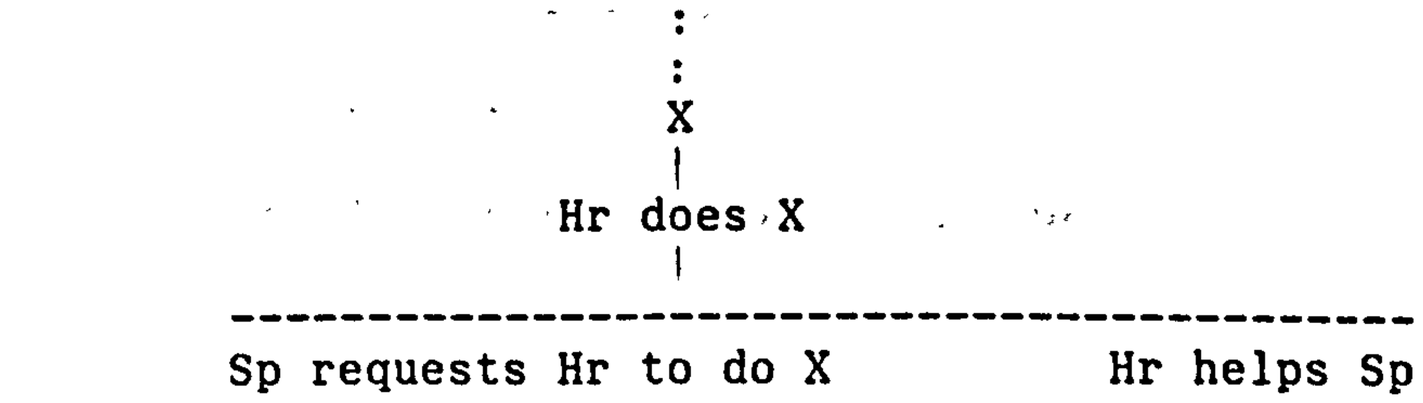
7.3.2.3. A degenerate explanation of requests

I've drawn a distinction between the fact of requesting X, and the fact of wanting X. But because of what requests are, the fact that they were made can also easily be explained in rather the same way that one would explain the fact of wanting X. Suppose that I offer you rules rather like this:

Hr does X <- Sp requests Hr to do X & Hr helps Sp

X <- Person does X

Then any plan that explains the action requested will also be able to explain the fact of the request. If the action X is explained as being part of a plan, then the plan can be extended like this:



But this is not the only sort of explanation available.

7.3.3. Why explain?

Why should Sp give explanations? I suggest there are four sorts of benefit he can get. One stems from the fact that just giving an explanation maybe worth doing; there is a social pressure to do so. The others, from the fact that after Sp has given an explanation, Hr has a fuller and more up-to-date version of Sp's plan before him. As a result,

- Sp may have made himself more cogent

- Sp may let Hr know better what Sp is about to do

- Sp may be able to forestall wrong interpretations of what he is doing

7.3.3.1. Sp may have discharged social obligations to explain himself

It is rude to leave your companions in too much doubt about what you are doing.

(From a story) A man and a woman have just murdered their employer

Woman: Now lets go to the office.

Man: Why to the office?

Woman: That's the nearest place we can get a typewriter with a carbon ribbon

Neither the reader nor the man can guess why a typewriter is wanted. I hope you agree that woman seems rude. Part of the woman's rudeness is due to her having refused a request for information. Any refusal may be rude. But part of it is because she leaves her accomplice in the dark about her intentions, when he has a social right to know about them. I can only speculate why this social pressure exists. A possible evolutionary reason would be this. Any group of agents who usually co-operate are going to do better if they keep each other up to date on what they intend, since if they do, each of them can both keep out of his companions' way, and advise them on problems with their intentions that they haven't yet seen. If you're among friends, gratuitous frankness may pay. In any case not to be told what other people are doing can be very annoying, and the specific remedy may be to explain what you are doing.

Here is another example.

A and B have just landed at Linate airport for the first time. Neither is sure where to go. They are now standing waiting for a bus. Suddenly A walks away

A: I'm just going to put this <apple core> in a bin.

A is of course explaining the action that he is doing. He has outlined his plan, and the action that B can observe fits into it. If he hadn't, B might have explained it by supposing that A was wandering off on some private travelling scheme that may separate them. This would alarm B. It is in order to prevent that that A says what he is really doing.

7.3.3.2. Sp may have made himself more cogent

7.3.3.2.1. Cogency of value

Suppose I make a request of you. If your have the sort of relationship with me that lets me make requests, you will probably feel warmly enough towards me to carry it out, for no better reason that that I ask. But if you're just doing it because I ask, there will be no reason why you should feel any particular sense of urgency about it. You only feel a request is urgent when you've got a reason for thinking it is. Some requests show their urgency on their face. For instance, "Get this dog off me". Others are taken to be urgent because they are only likely to be made in a serious situation. For instance, "Ring the doctor".

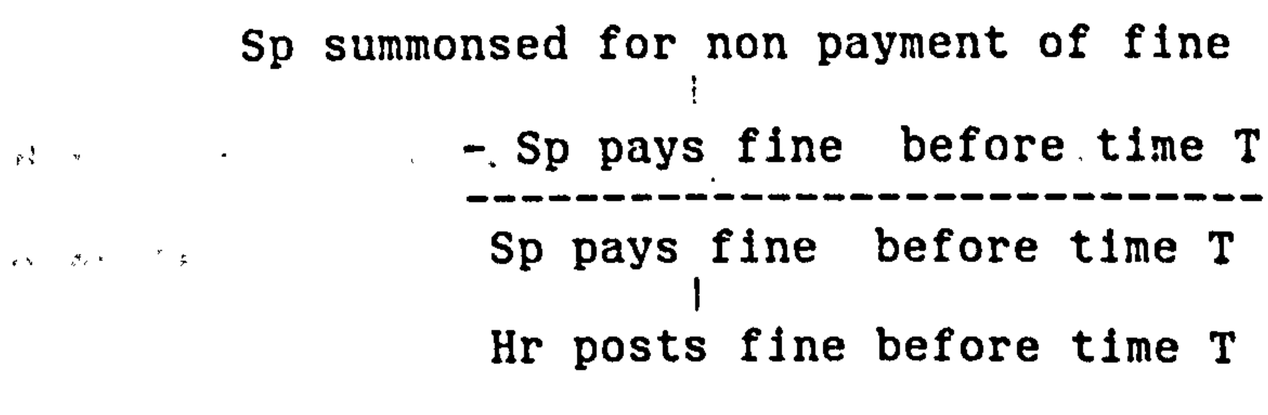
But if you do want me to jump to it, you're going to have to make it obvious that the request matters. How can you do this? You have a choice.

- You can label your request as urgent, and expect me to trust you about that. For instance, you can say "Lend me a quid. It's urgent!" or "Please be absolutely sure not to forget the eggs". I won't say more about this option.

- You can demonstrate that you request is urgent, because complying

with it procures some especially good goal, or not complying causes some particularly bad problem. The way you do that is to let me see what course of action (which may be plan of yours) the result of my action is going to assist.

For instance, you say "Will you post this letter for me?". I agree, and put it in my pocket, and as like as not forget it for a couple of days. If however you say "Will you post this letter for me? It's my parking fine and if I don't pay it I'll be summonsed", you explain the substance of your request by showing how it is embedded in a plan like this.



Indeed, just saying "Will you post this? It's my parking fine" would be enough to allow the reconstructioin of this plan. What matters is that the anti-goal, being summonsed, is now clearly seen as the effect of non-compliance, and the importance of getting the letter off is correspondingly enhanced.

(A minor point. I have been using "urgent" as ambiguous betwqeen "urgent" in the sense that the opportunity for doing something must be seized rapidly, and "important", meaning that the goal or anti-goal is something that you care about very much. But revealing a goal that is either of these will have the same effect.)

7.3.3.2.2. Cogency of fact

Just as Sp may want to increase the likelihood of Hr becoming really interested in fulfilling Sp's expressed wants, so may he feel he has

to strengthen the chance of Hr believing his expressed beliefs. Again he has two ways of doing this.

- He can label his remark as especilly sincere and persuasive - "You must believe me, I'm telling you the truth".

- He can demonstrate its truth by showing the grounds he has for believing it. One can make a distinction between giving evidence that something is or was true, by showing that it occurs in an executed plan, and showing that it will be true, because it occurs as an effect of a plan yet to be executed.

Here is an example of a plan explaining a fact by means of an executed plan.

A: You'll find the book on the hall table. I left it there this morning.

And one of a future fact.

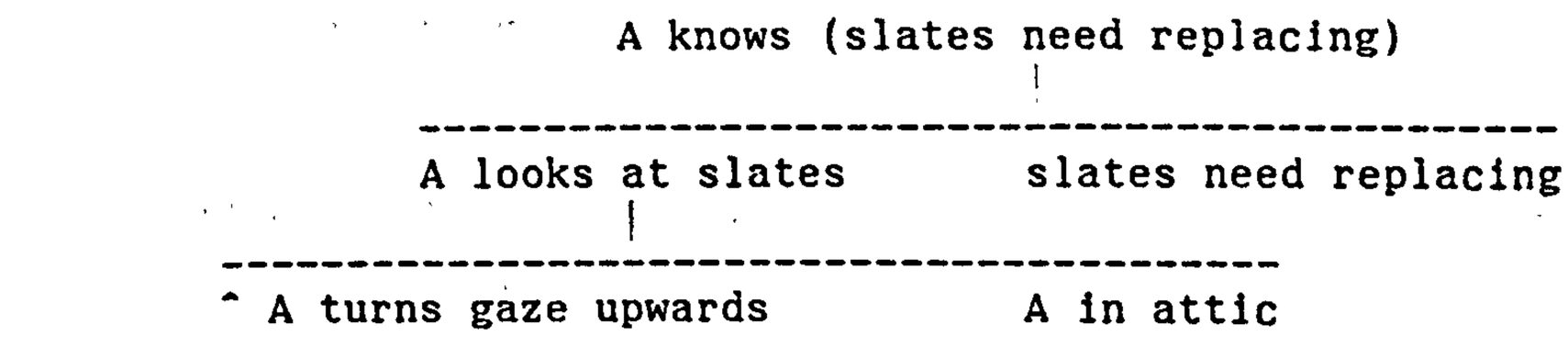
A: I won't be home when you get in. I'm going to a CND meeting.

In each case, A is involving his main remark in a plan of which he must, since he is the agent, have special knowledge. Since what he says is good evidence for what he himself does or will do, it must be equally good evidence for any part of what he does.

An explanation of how a fact came to be known may also increase the cogency of a statement. For instance:

A: Some of our slates need replacing. I was up in the attic yesterday.

The executed plan behind this is something like



Note that if the example had been "The ballcock is broken. I looked in the tank." then one could object that one didn't have to consider A's plans. One might just say that there was a direct inference, about A's authority on the matter of broken ballcocks, like this

A is authoritative on whether (the ballcock is broken) <-

A looks in the tank

This is a rule one could justify by considerations of plumbing and vision. But in the slates example, does one want a rule like

A is authoritative on whether (the slates need replacing) <-

A was in the attic

No. It would be true only if, if A looked up while in the attic, then he would see the slates. That is not a condition of the rule. And the idea of a rule like

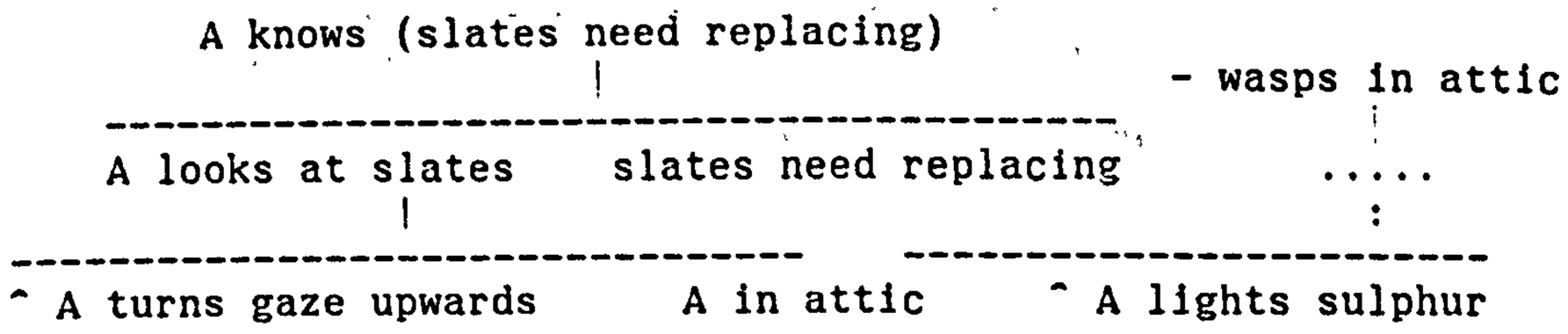
A is authoritative on whether (the slates need replacing) <-

A smoked out wasps

to explain an example like

A: Some of the slates need replacing. I was smoking out the wasps yesterday.

is laughable. On the other hand, this (rather abbreviated) plan with its side-effect would explain the connection.



7.3.3.3. Sp may let Hr know better what Sp is about to do

All the usual reasons for Sp letting Hr know what Sp is doing may apply - unspoken communication of fact and value, giving opportunity for criticism and comment, and so on.

7.3.3.4. Sp may be able to forestall wrong interpretations of what he is doing

Sometimes one can offer an explanation of an action, not just to cure the alarm caused by the person observing you not knowing what you are doing, but to get rid of some specific plan that they wrongly think you are following. For instance

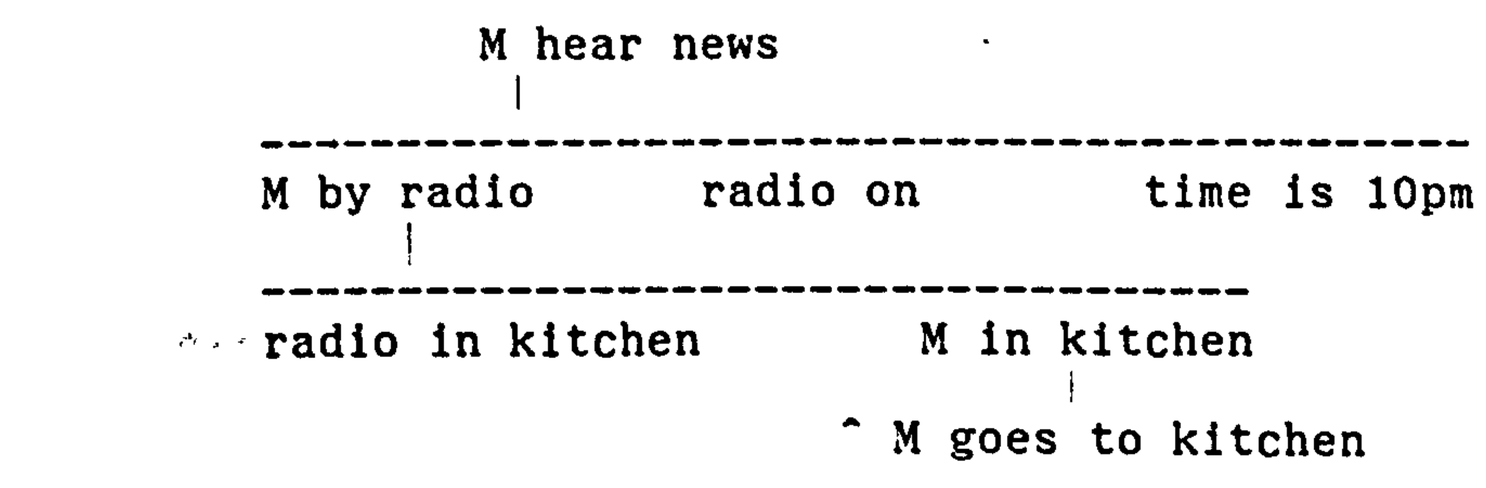
example 5/17

M usually tries to listen to the 10pm news. The radio is in the kitchen. He walks out of the sitting room where he and W have been sitting

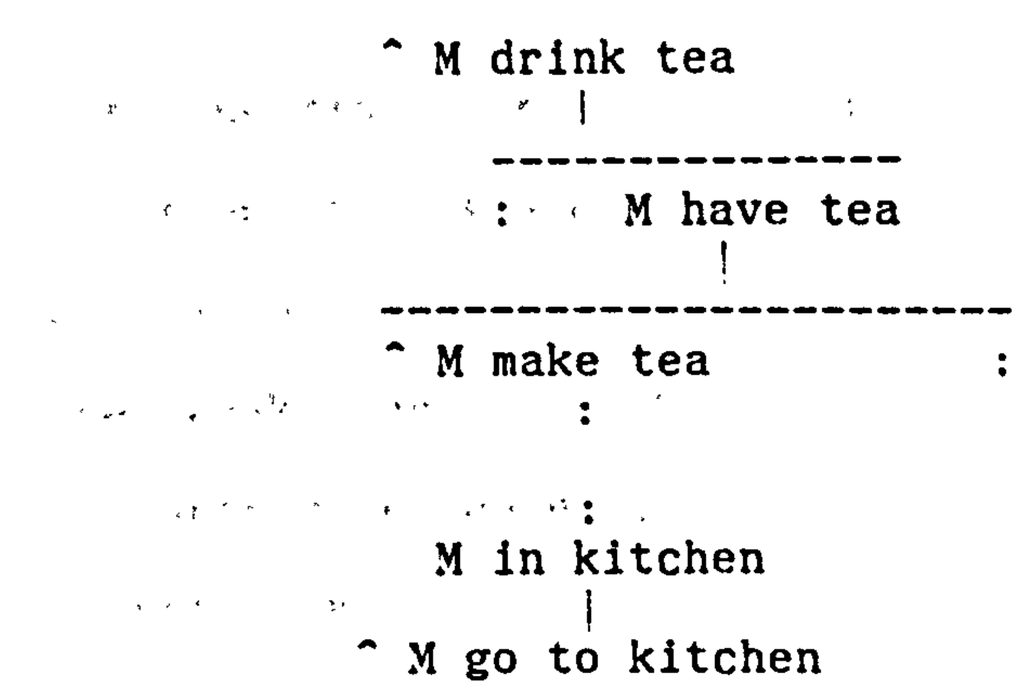
W: It's 9pm, not 10pm.

M: I'm going to get some tea.

What happened? M left. W guessed that he had this plan



W, who seeks to help M, tells him that a precondition of this plan, that it be 10pm, is false, and so his plan is unsound. M rejects the warning by telling her that in fact he has a different plan, and isn't injured by what she has told him. He does this by revealing that his plan (in outline) is



Why should M have made this specific rejection of W's advice? Could he not just have ignored it? A possible reason for being explicit about the rejection is this. Criticism of your plans by your friends is useful. They know this. But if they offer you criticism which you reject without reason, they will come to think that you don't value it, and may eventually stop giving it to you. To prevent them giving up, one feels an urge to show why one doesn't accept the particular advice they have offered, without denigrating their advice in general. Indeed, giving reasons for one's rejection is not in fact

getting good advice and then explaining why you don't use it, but instead it is showing that it isn't advice, and that your plan isn't the shape that their advice presumes it is.

Here is another, slightly more complex, example.

example 5/22

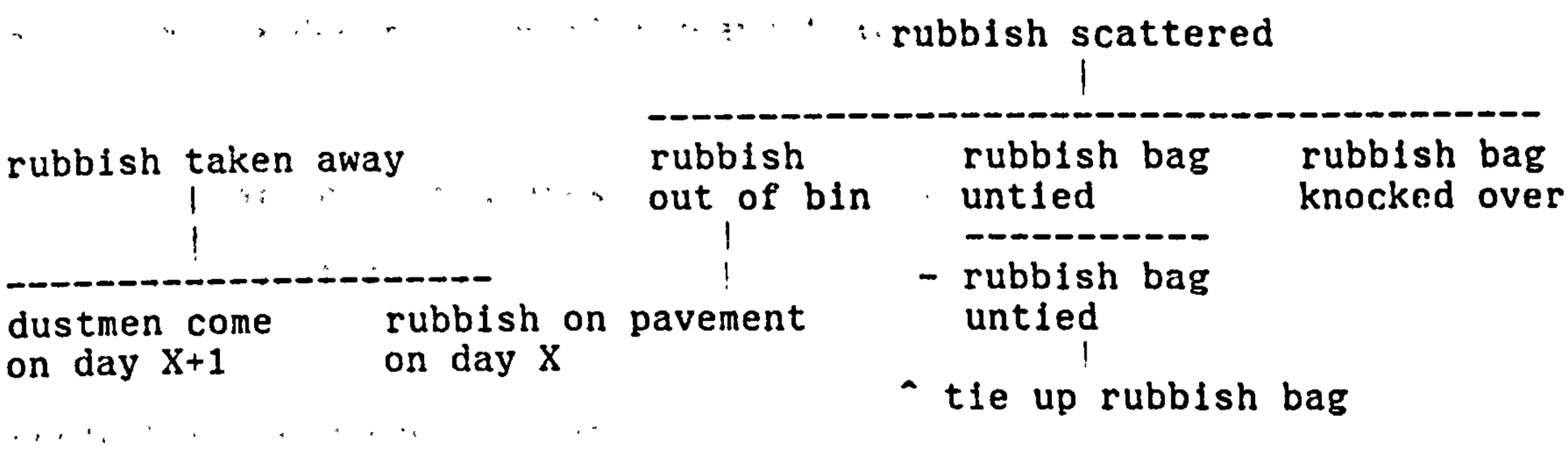
M is tying up the neck of a rubbish bag which lines the dust bin, which is usually done the day before the bag is put out for the dustmen

W: Is it rubbish day tomorrow?

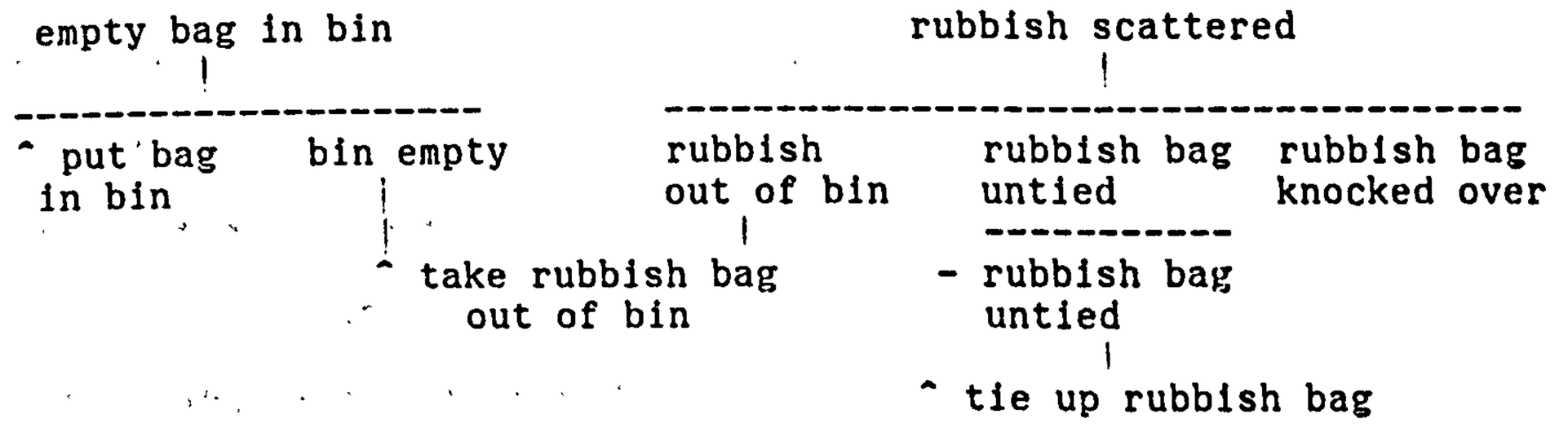
M: No. The bag is full.

W's question can be interpreted in several ways; as a rhetorical question to dissuade M from a mistaken activity; as an attempt to get an explanation of what he is doing; as a question about the day of the week, prompted by what she has just seen M do. But in any case M's second sentence is telling her that she is on the wrong lines. He is not tying up the bag so that he can put it out without the rubbish blowing about, but because it is full and a new one must be put in the dust bin.

The plan that W at first contemplates M having is the plan below. She believes M is attempting to prevent an unpleasant possible side-effect of a particular plan. "Rubbish scattered" is bad. "Rubbish taken away" is good.



M proffers as an alternative explanation an identical effort to avoid a bad side-effect, but it would arise from a different plan.



7.3.4. Sequence of explanatory utterances and explanands

In most of the examples so far, when the explanandum has been an utterance, the explanandum has come first, followed by some other information that lets one build the explanation. But things don't have to happen like that.

> A: I'm going to Bauermeister's <a bookseller>. I want a Spanish dictionary.

could have occurred as

> A: I want a Spanish dictionary. I'm going to Bauermeister's <a bookseller>.

in which case the second sentence is not intuitively an explanation.

but rather a description of a means that will be pursued. Similarly,

> A: Some of the slates need replacing. I was smoking out the wasps yesterday.

could have been produced as

> A: I was smoking out the wasps yesterday. Some of the slates need replacing.

My intuition is that you can suppose that these reflect different intentions. The first is probably said by someone who is interested in getting consent to his suggestion that something ought to be done about the slates. The second could have come from someone who while describing what he did yesterday was more or less idly struck by a thought that something would have to be done about the slates; but it could also have come from an avid slatemender who wanted to lay out the strength of his evidence first.

The flow of thought of the man who reflects about what he saw in the attic is not for me. I only attempt to explain purposive speech. Taken as purposive speech, it seems to me that the order of the sentences produced by the slatemender doesn't matter. What he was after was not explanation as an end in itself, but in order to give his important remark the added cogency of being seen to have been verified by what he had done. His remarks add up to the same thing in both cases.

There may well be a reason that he could give for choosing one form over the other, but I believe it would have to in terms of, for example, presenting facts in the order that made an inference easier, not in terms of inducing in his hearer one inference rather than

another. It would be a decision made on grounds of rhetoric rather than logic. I expect it would be rational on such grounds, but I have no idea what the criteria in that region would be.

If this is so, one may have to wait over several sentences before one is able to see their intended relation to each other. But as my examples show, I have only looked at exchanges of a very few sentences, and can say very little about longer passages.

7.4. Conclusion

Some discourse events are constituted by changes in the plans that a planner is known to have. I have given examples. Either to identify or to respond to them, one is going to have to keep a model of what the plans of the other parties to the discourse are. This is made easier if such plans do not have to be attached to particular people but can be taken as global to the discourse, in the way suggested in the previous chapter.

Chapter 8. Attempts at processes

This chapter is a sketch of two processes that could detect the changes, and the point of the changes, being attempted by the maker of an utterance.

The first approach was to ask, Is there a fixed case analysis, in the form of a flow-chart of questions, that one can follow to find the purpose of a remark? Understanding an utterance is finding a path through the analysis compatible with the utterance and the currently extant plans. I show some such flow-charts. For simple cases they are appealing, but in general they fail: the purposes of utterances are too various.

The second uses chains of inferences built from rules whose antecedents are purposely not checked. If the chains that can be built starting from the utterance and an extant plan run to a common "interesting" fact, it may be wise to see if these chain can be made into proofs. If they can, a wise planner will make certain changes to his plans. These changes may be credible purposes of the utterance. This approach is I suspect the correct one. However it runs into practical problems of search control.

Though these approaches are both rejected, they are only means of realizing the theory. Neither failure impugns the theory itself.

8.1. A speaker's remarks can affect his own plan

What follows invokes the notion of what A says altering his own plan. This sounds a little odd. In fact whatever A has planned to do remains the same whatever he says. However, what matters here to A

and his interlocutor are the plans they mutually believe themselves to have. A is capable of altering this by what he says; for instance. he does so if he reveals a previously private intention; or if he asserts some fact not previously mutually known but which makes some plan he is believed to intend impossible.

The image I had in mind while trying to form these ideas was this: A and B are sitting at a table jointly doing something such as carving some sculptures. They also each have their own private sculptures out of each others' sight. The joint sculptures are supposed to be like their private sculptures, but at each moment the public sculpture may be merely a partial or even a wrong version of the private ones. They can chisel at the joint sculpture or at their own, but not at each other's. The joint sculpture is the mutually known plans, and can be affected by what each of them allows it to be deduced that his prvate plan (or sculpture) looks like. If A (say) alters the public sculpture and then says that his private sculpture looks like the new one, then what he is doing in analogous to what I have described as "A altering his own plan by what he says".

8.2. Is a fixed flow-chart of questions possible?

Once I had realized that the alterations a speaker revealed in his own plan might in themselves have the effect of an utterance, I wanted to see how they worked. I argued like this. If these unspoken utterances are really like ordinary utterances, then they should have the same range of effects as ordinary utterances, and so whatever mechanism was used for understanding direct remarks should be more or less applicable to unspoken ones.

Furthermore, what a speaker says can only effect two classes of plan: his own and his hearer's. There is (usually) no one else listening.

And if it affects his own, so as to give rise to an unspoken remark, that unspoken remark can only be expected to matter to his hearer's plans. To see this, consider what would happen if this weren't true. The idea is that an unspoken remark that a person generated by offering a contrast between two versions of a plan he has feeds back to affect another plan he is known to have.

Sp says F

F alters his plan P1 to P2

P1 when contrasted with P2 reveals that H

H alters another of his plans Q1 to Q2

Q1 when contrasted with Q2 reveals that J

J has some useful effect on Hr's plans

But why should Sp go about communicating J this way? Surely it would always be to his advantage to leave out one stage and just say H explicitly? This argument may not be very compelling: it seems to be an argument against any unspoken utterance, since one could ask "Why didn't Sp say what he said with one indirection with none at all?". On the other hand, it may be the case that even though one indirection works, two indirections may be too complex for Sp to be able to be confident that he would get his message across. And I have come across no example that seem to need double indirection.

It seemed to me then that what I could do was draw a net, and label the arcs with events that could occur during the chain of events that led from an utterance to its having its having intended effect. If this net were exhaustive, and if one could at each arc test to see whether the event it was labelled with could have occurred, then one could search throughout the net. If one came out, it would mean that there was a course of events that led to a benefit for Sp, and which might be the purpose of his remark.

The net would be made of several parts that would be repeated in more than one place. Such parts would be tests to see whether some particular remark could alter a particular plan; to see what information contrasting two versions of a plan revealed; and to see whether people tended to help each other and would therefore benefit from the bits of information being passed round.

Let me start with statements. What sequences of events ought to be recognized in the net? First of all, some statement has to be made before anything can happen at all.

A statement, I thought, can then do one of two things to a plan. It can describe the plan; or it can alter it, by stating that the world is different from how the plan takes it to be. There are two people who may have plans involved in the exchange; the Sp and the Hr. So the effect of a statement must fall into one of the cells in the grid:

	alter plan	describe plan
Sp		
Hr		

No action of Sp can fill the cell of "describing Hr's plan", because he doesn't know what it is. (I know he can purport to do this; "You want me to stay in the rain so that I die and you will become master of Headlong Hall! But you shan't, you shan't". But this cannot actually be description of the Hr's plan. At most it is description of what Sp takes it to be.) The other cells can be filled.

The simplest operation and the most obvious sort of benefit this can

procure is the altruistic one of advising the hearer. This falls in the cell of altering Hr's plan. However, this is only any good to Sp if he wants to help Hr.

This leaves the cells where Sp is describing or altering his own plan. Both of these will have their effect through Sp showing that his plan is different to how it was thought to be, so that new information arises from contrasting them.

As far as describing his own plan goes, it seemed that there were two ways this could be done.

- Sp can more or less explicitly describe his plan. He can say "I'm going next door to borrow a cup of sugar".

- Sp can just reveal his need, which may be enough evidence to let Hr guess what plan he will pursue to that end. He can say "I want a pint", and leave Hr to infer that he will go to the pub to get it.

Sp can alter his own plan in just the way that he can alter someone else's. But he can also say something with the intention, not of altering a plan that he is known to have, but of being credited with a previously unknown plan which must be imputed to him before his remark makes sense. For instance, a man announces he has no money. If you want to assume that this alters his own plan, but you don't know what plan he has, you have to assume he has some plan that it does alter. In this case, a plan to buy something would meet the bill. This I call "forced ascription", because one is forced to ascribe a plan to the Sp. When this has occurred, the contrast of no plan and the revealed plan may affect Hr.

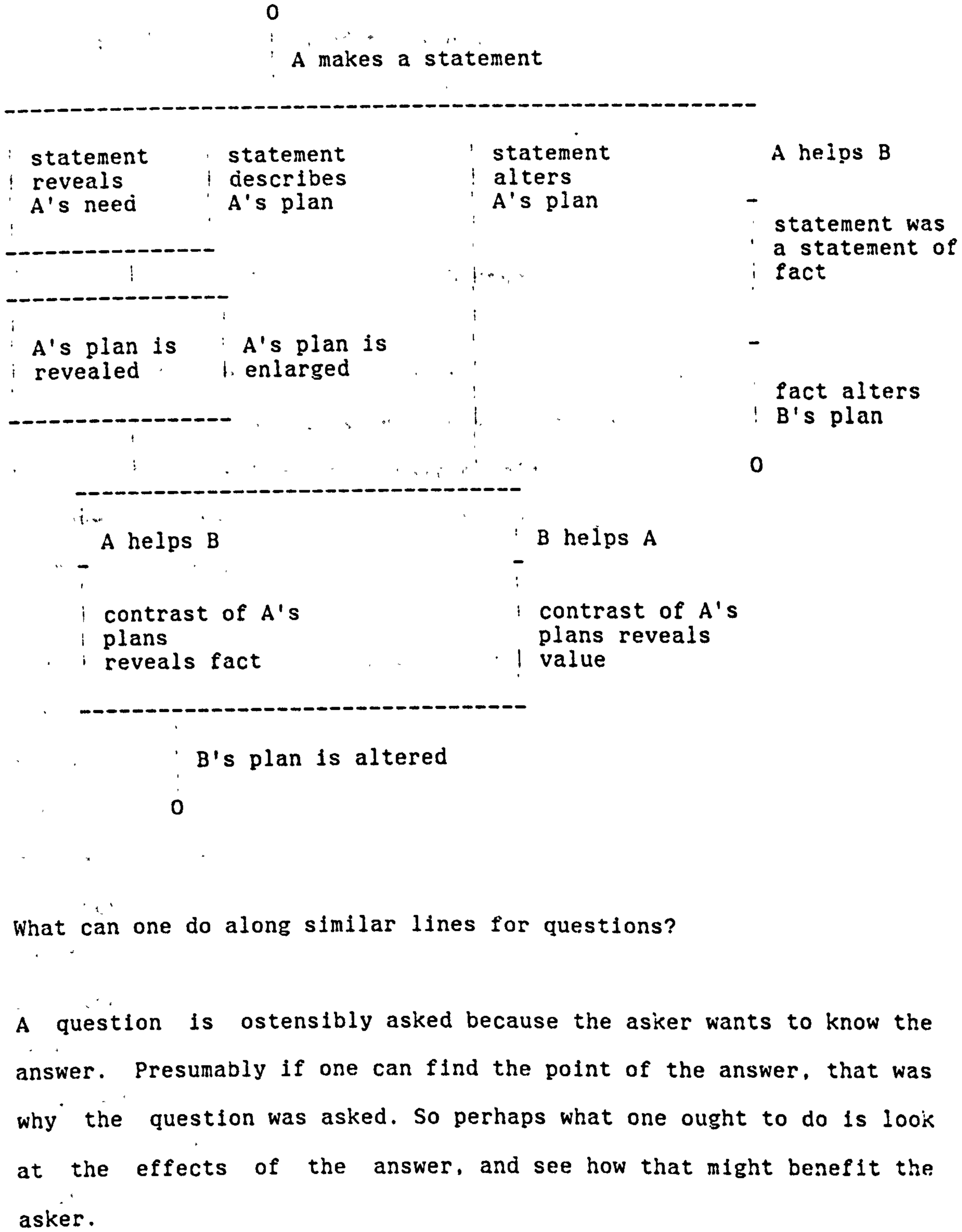
But after either of these, it still has to be shown that the

information derived by contrasting plans has an effect that Sp finds useful. Two sorts of information can emerge from the contrast:

- Information about what will happen, about fact. This may make Hr change his plans. This will only please Sp if Sp wants to help Hr.

- Information about what Sp wants, about value. This will only have any effect on Hr's plan if Hr wants to help Sp. Otherwise Hr will just ignore Sp's wants.

In the light of that, here is the net I devised:



What can one do along similar lines for questions?

A question is ostensibly asked because the asker wants to know the answer. Presumably if one can find the point of the answer, that was why the question was asked. So perhaps what one ought to do is look at the effects of the answer, and see how that might benefit the asker.

Rather than talking about Sp and Hr, which leads to confusion, talk about two people A and B. A asks the question. The net will start with B saying something in reply to A.

The grid of possible operations is the same, except that it should be

labelled with the names of the speakers, thus

	alter plan	describe plan
A		
B		

The cell that can't be filled is B describing A's plan.

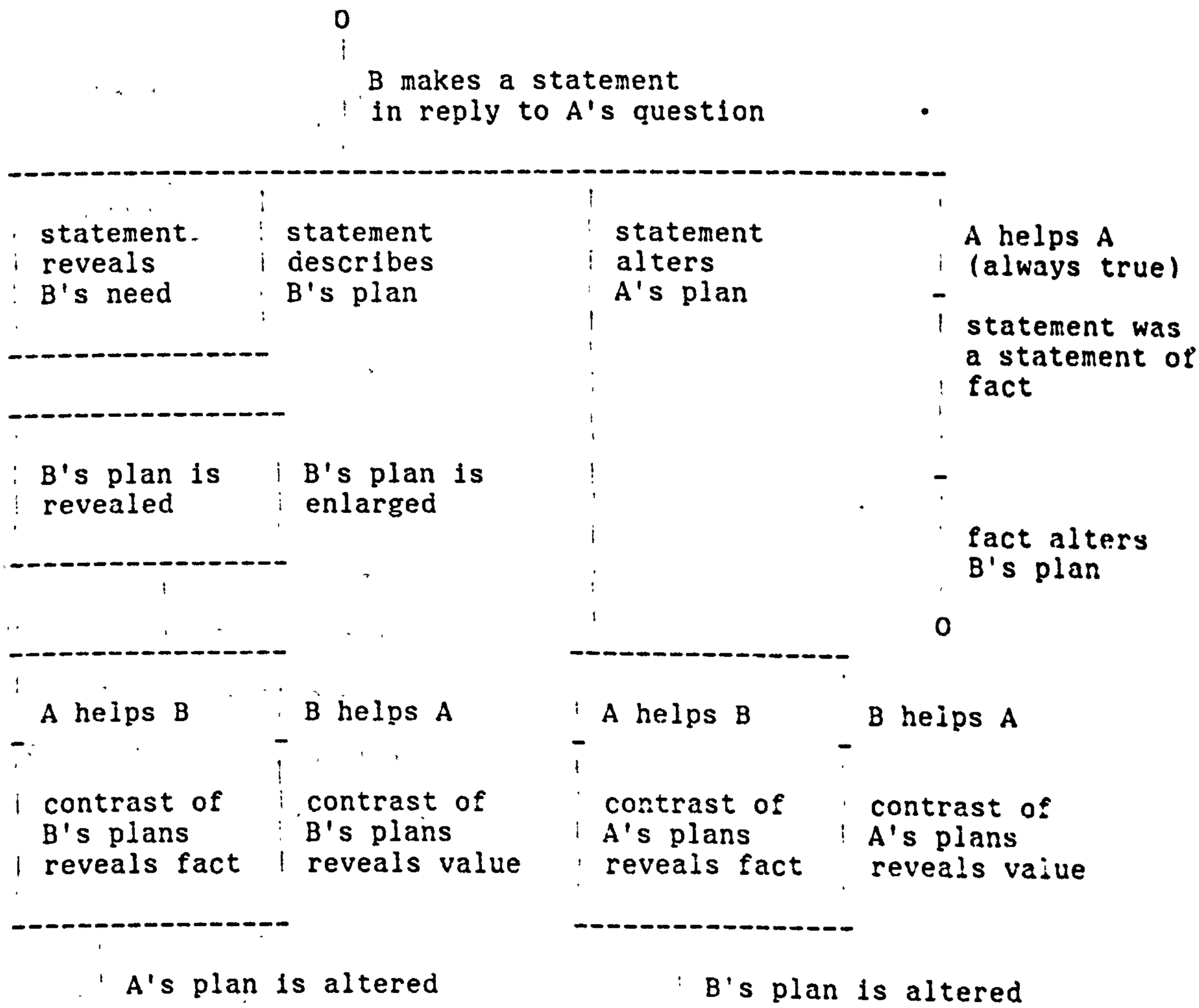
The simplest benefit that can arise is for the answer to the question to affect A's plan, as when A asks if the last bus has gone. What A wants is the effect of the answer on his own plan. That is a sequence of events comparable to what happens when Sp makes a statement altruistically advising Hr.

Comparable to a statement where Sp describes his own plan to give Hr information would be A asking B a question about B's plan, in order to extract information from it. A question asking B about his plan ("What are you doing with that cup?") or about his goals ("Do you want a cup of tea?") will do this. If B answers, what he says will be a description of his plan. A can perhaps derive information from the contrast of the plan he thought B had, with what he now thinks B has. But then, just as before, this information has to be proved to be beneficial. If it is fact, it is good if it alters A's plan in any way, since it brings it into better accord with reality. If it is value, it only benefits A if A tends to help B, and is therefore glad to learn of services he can do.

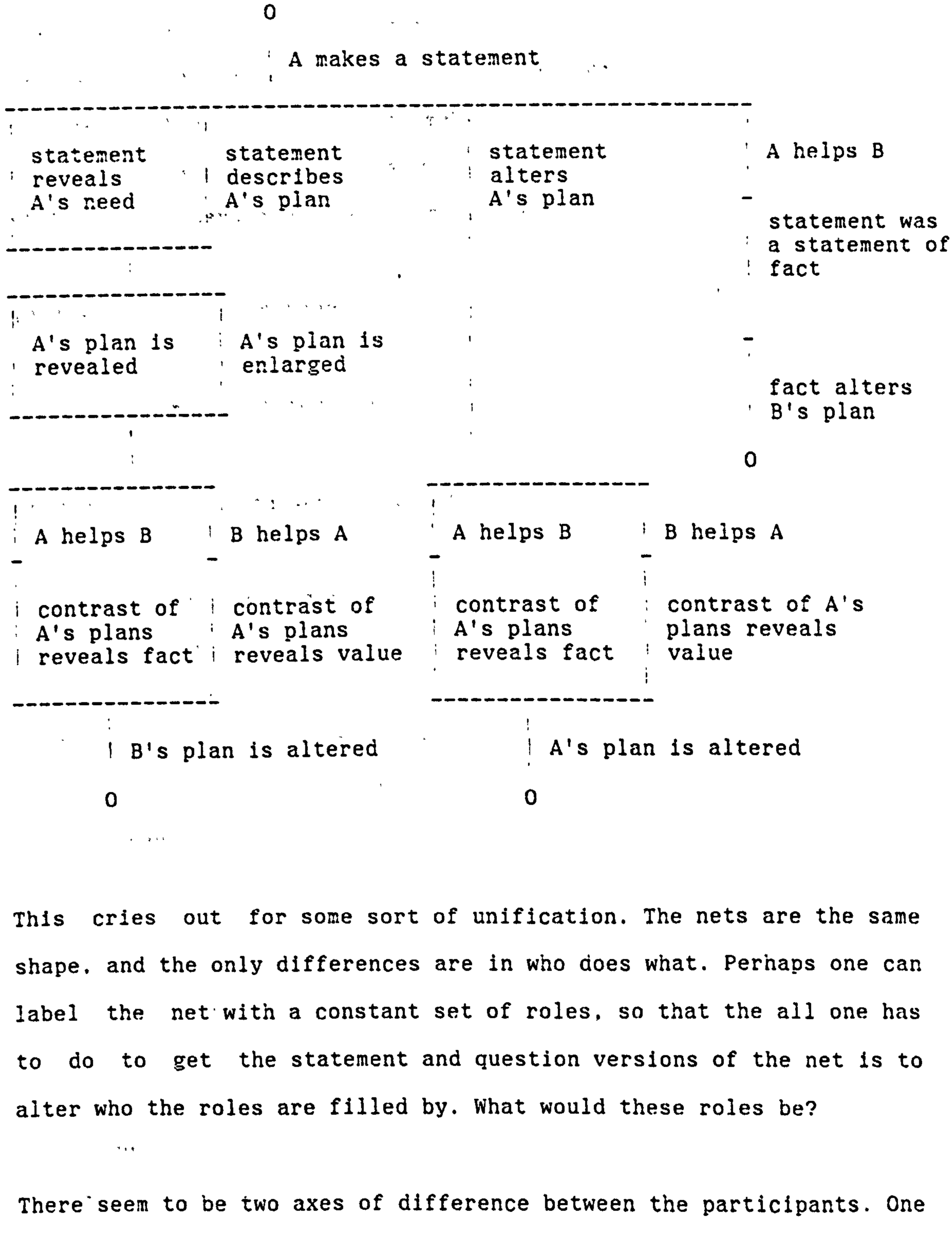
Comparable to Sp saying something that forces Hr to ascribe a new plan to Sp is the case of A asking something to which the answer will alter, not a plan he is known to have, but a previously unknown plan. Then the contrast of no plan and the revealed plan may affect B. As far as guessing which plan is to be ascribed, any method that works when fed a statement that Sp made to Hr will also work when fed an

answer that that B may make to A. Of course the answer may be "yes" or "no", so two cases may have to be tried. But I stress that, in either case, the mechanism would still be working from a statement, even though a statement that hasn't yet been made. The information derived from the contrast will benefit A if it is fact and it helps B when A tends to help B, or if it is value and affects B because B tends to help A.

The net for questions then looks like this:



These nets look very similar. They can be made even more similar if one notices that the first can be unfolded, that two identical paths can be written out to duplicate a single one. If one does this, one gets



This cries out for some sort of unification. The nets are the same shape, and the only differences are in who does what. Perhaps one can label the net with a constant set of roles, so that the all one has to do to get the statement and question versions of the net is to alter who the roles are filled by. What would these roles be?

There seem to be two axes of difference between the participants. One axis is about direction of flow of assertion. There is always an utterer and a receiver of the remark that matters, that is, the remark that has an effect on the participants plans. The remark that matters is the statement in the case of the statement, and the reply in the case of a question.

The other axis is about who causes the exchange. The person who causes the exchange is the one whose motive must be examined. In a statement, the causer is the maker of the statement. In a question and answer exchange, the causer is the person who asks the question. I'll call the person who isn't the causer, the other.

In a statement, this happens:

A: (makes statement)

In a question and answer, this happens:

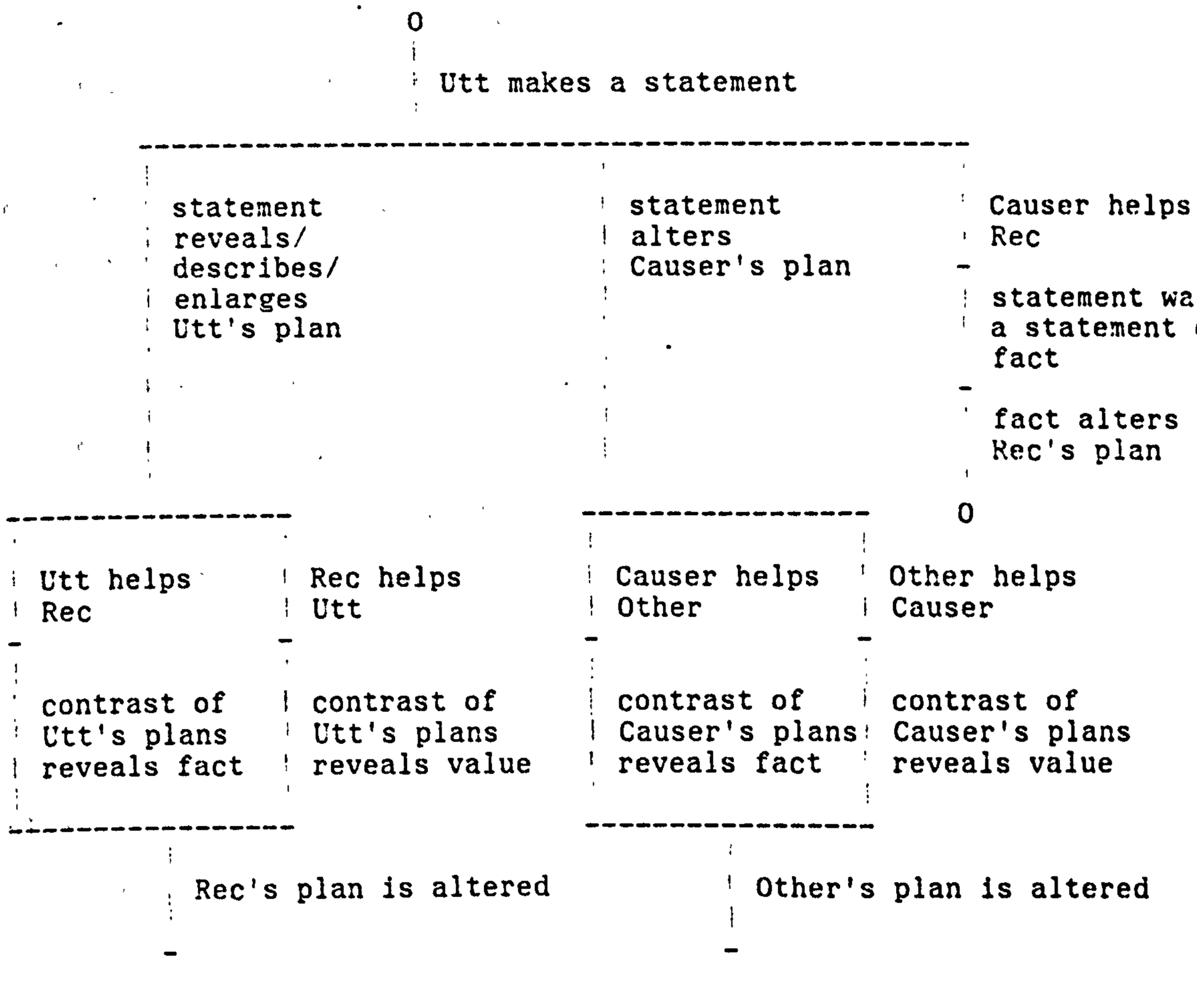
A: (asks question)

B: (makes reply)

With the actors as here described, the roles are filled like this:

	statement	question
causer	A	A
other	B	B
utterer	A	B
receiver	B	A

What does the net look like with constant roles? (I've abbreviated the top left hand corner.)



I abandoned this approach because it seemed too artificial and too constricting. In particular, there is no obvious place in it where machinery to recognize offers, refusals, etc can be inserted.

8.3. Looking for the point of an utterance using chains of inference

My next attempt was to try and understand an utterance by looking for some beneficial effect it had on some extant plan, updating the plan, and repeating the process for the next utterance. In slightly more detail,

- The system believes that there is a set of plans that each of the speakers is either contemplating or executing, and that what these plans are is mutually known. The sytem knows that it may have very imperfect knowledge about the plans. Some of them may be wholly

unknown. But whatever the state of knowledge, it is believed to be mutually known. It is possible that the parties actually have private beliefs and intentions as well, which may differ from what is mutually known. This will occur during attempted deception. But I am going to ignore that.

- A speaker makes an utterance. The system has to see how this is beneficial. This involves
- seeing the effect of the utterance on the plan
- seeing that it is beneficial

- Once the system has discovered the effect on the plan, it can suppose that the change is mutually known. Why? Because the initial plans were ex hypothesi mutually known; all the speakers are assumed to apply the same mutually known recognition process; if this involves postulating that a plan has been changed, then the changed plans must be mutually known too.

- Now the whole process can be repeated with the new plans.

Seeing the effect and seeing that is beneficial each have their own complexities.

Remember that the system knows that it may not have a complete picture of the other persons' plans. It is possible that the utterance should be intended to have an effect on one of the unknown parts of the plan. To detect this, the system will first have to guess the unknown parts of the plan. Having to do this won't be uncommon.

A: I'm going to get some milk.

B: The shop is shut.

All that A reveals of his plan is its (presumably final) goal. B's response can only be understood if one imagines him having made a guess at the means that he thinks A will employ.

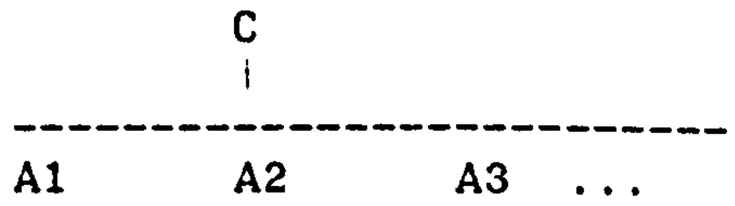
The same thing happens when the bit of the plan that is unknown is above rather than below what is known. In

A: I'll drop in at the bookshop.

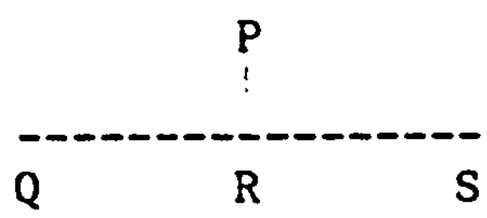
B: I don't think your order will have come in yet.

B must first have guessed at the purpose of A's plan, and then attacked the new bit.

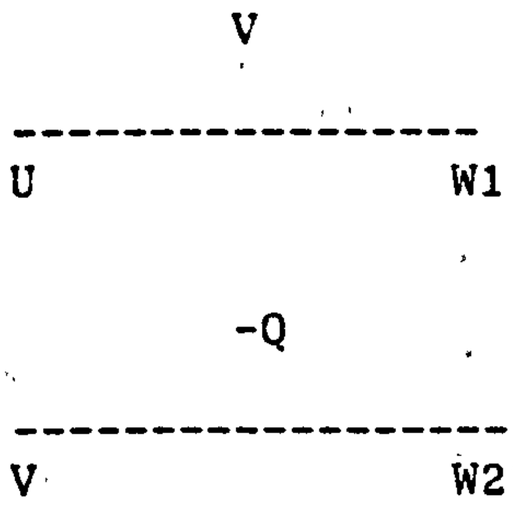
How can the system guess at the unknown bits? My approach assumed that there is a set of rules all of the form



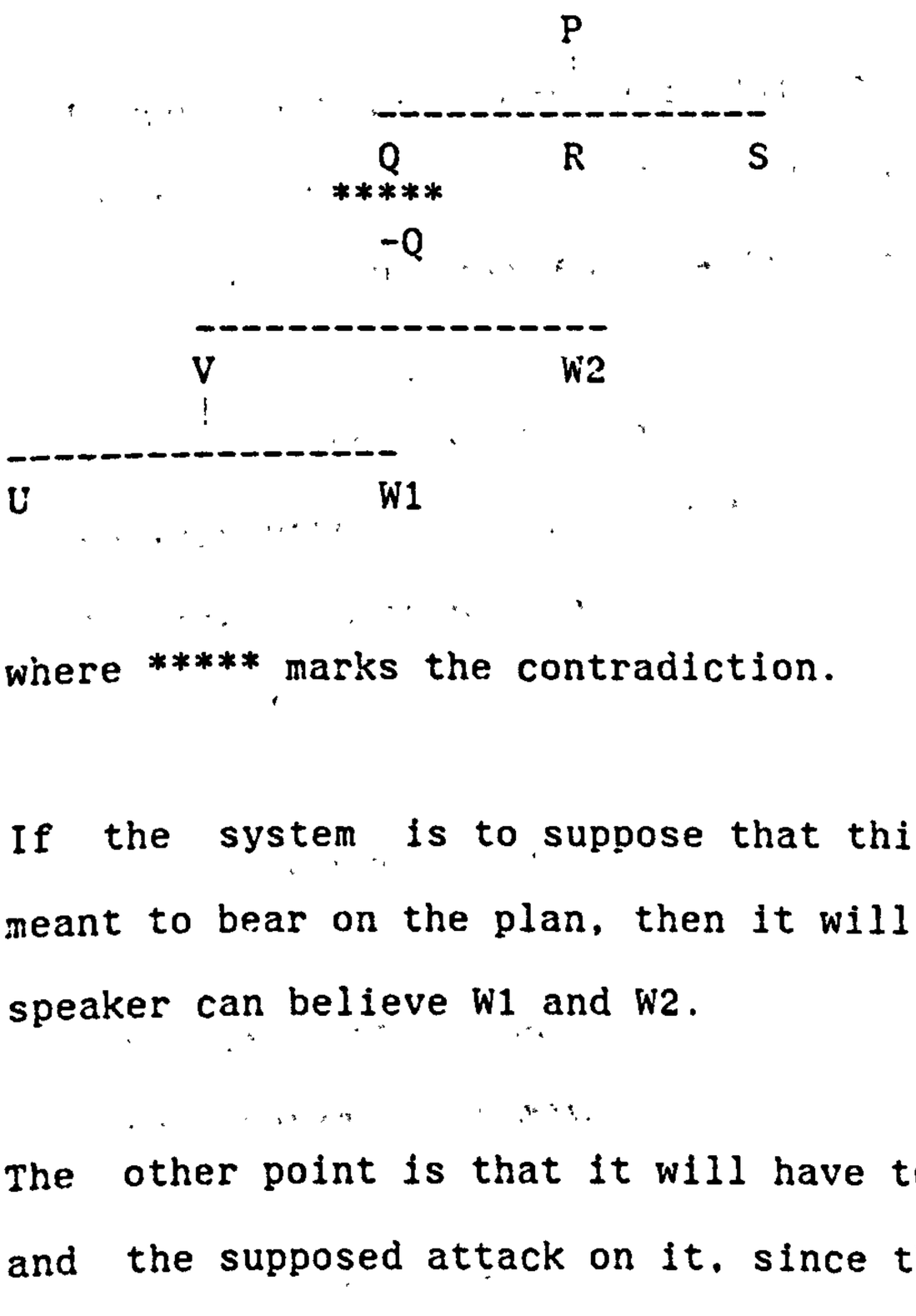
Now suppose that the plan attributed to some speaker is



and the utterance that someone makes is "U". The system can search forward from U using the rules it knows until it finds a possible conclusion that would affect the plan by contradicting one of its preconditions. The action of "searching forwards from a sentence S" is just finding a rule of which S is an antecedent, finding the consequent C, and then perhaps repeating the procedure with C. Or it can "searching backwards", which is the obvious complement. For instance, the system might have the rules



and string them together so that it could conclude -Q. But -Q denies a precondition of the plan, thus



where ***** marks the contradiction.

If the system is to suppose that this is the way the utterance was meant to bear on the plan, then it will also have to suppose that the speaker can believe W1 and W2.

The other point is that it will have to record both the initial plan and the supposed attack on it, since that attack will itself be open to attack later, as in

A: I'm going to the shop.

B: It's shut.

A: It was open when I went past it a minute ago.

At the end of this exchange, the complex plan will something like this:

```
                    A buys ... from shop

                    ----------------------
                    shop open        ........
                    *****
B's remark          - shop open
                    *****
A's rejoinder       shop open

                    ------------------------------------------------------
                    A believes shop open    A authoritative about "shop open"
                                            ----------------------
                                            A observed "shop open"
```

I never worried about extracting such a proof from A's rejoinder - I would have pretended that A had said "A observed "shop open" " and tried to get the system to work forward from there.

Seeing a remark as an attack on an unstated goal of a plan needs forward branching, not just from the utterance but from the plan too. In the bookshop example above, all that A reveals of his plan is something like

```
        A goes to bookshop
```

If the system searches forward from this, and from the negation of the utterance, it may come to the same conclusion, postulating a plan like this:

```
        A has book
        ------------------------------------------------
        A at bookshop      A buys book      book at bookshop
                                                *****
        A goes to bookshop                      - book at bookshop
```

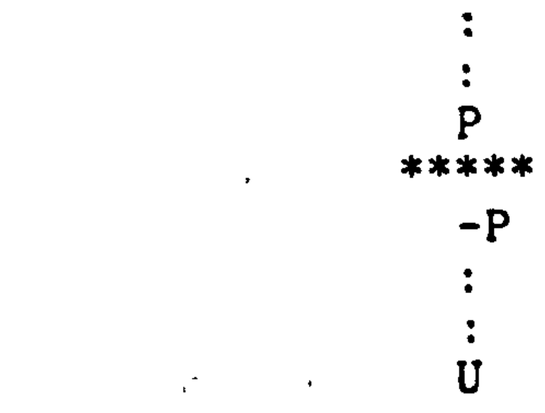
In this, what is mutually known of the plan has increased. If the system is going to use this as an explanation of what B said, then it is going to have to make this committment about what it thinks A's goal is. That is right. It lets one guess at what is happening in an exchange such as

A: I'll drop in at the bookshop.

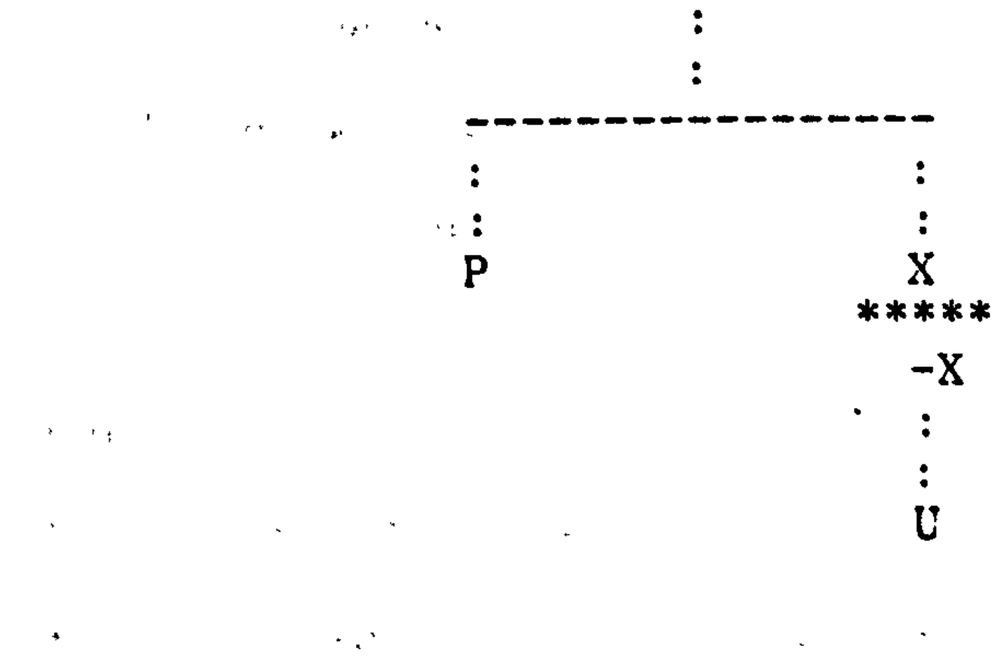
B: I don't think your order will have come in yet.

A: No, I wanted to get a pad of paper.

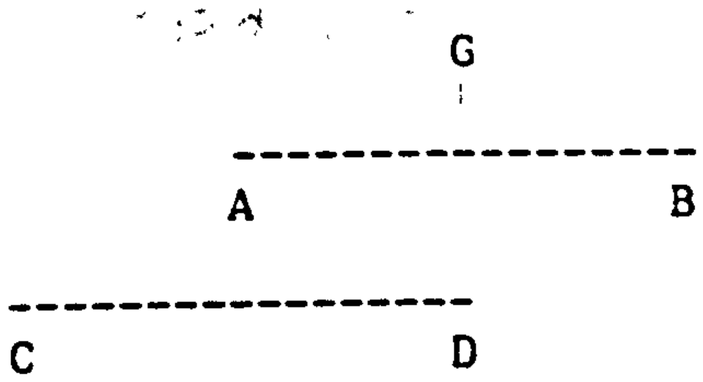
Looking for remarks to have this sort of relevance, attacks on the soundness of a plan is, it seems, in theory trivial. Suppose P is some node at the fringe of a known plan, and U is the utterance. The system must look for possible inferences so as to produce a proof tree in this shape:



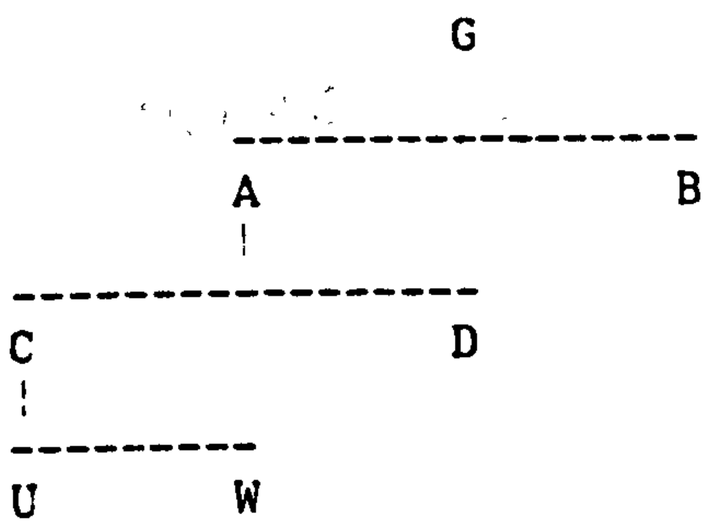
or in this shape, where it is guessed that if P is a precondition, it is because it together with other facts, such as X, are collectively the antecedents of some rule used in the plan:



An utterance need not impugn the antecedents of a proof tree to alter it. It may support it. For instance, it may be mutally known that one person has a plan, of which the mutually known part is



This does not mean that eg C is true. Perhaps it is known that it isn't. An utterance U could affect the plan by supporting C.

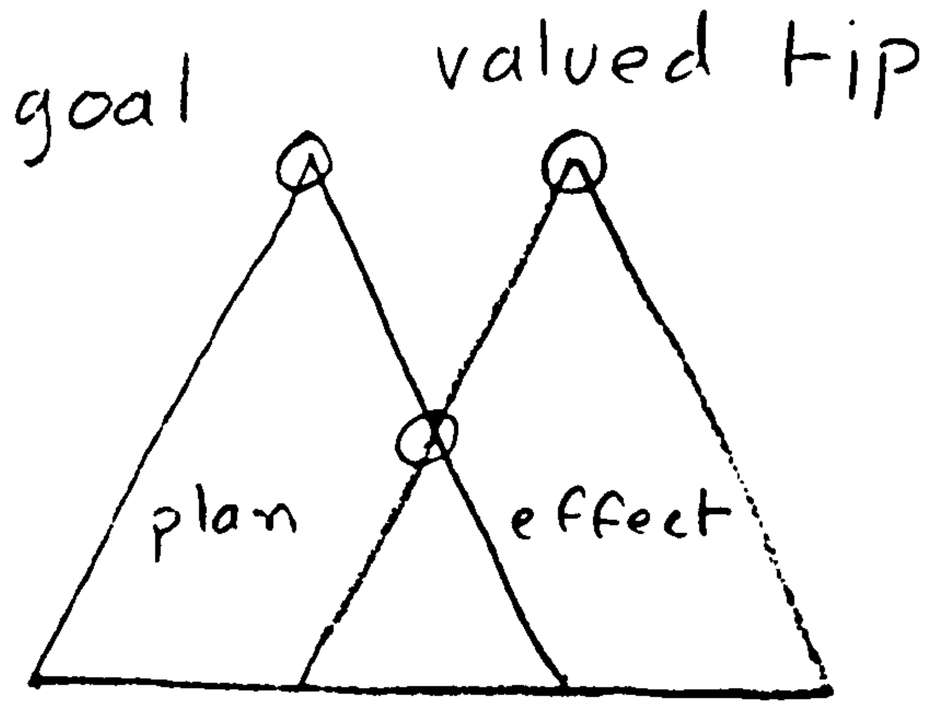


Discovering this needs search forward from U and backward from the known plan.

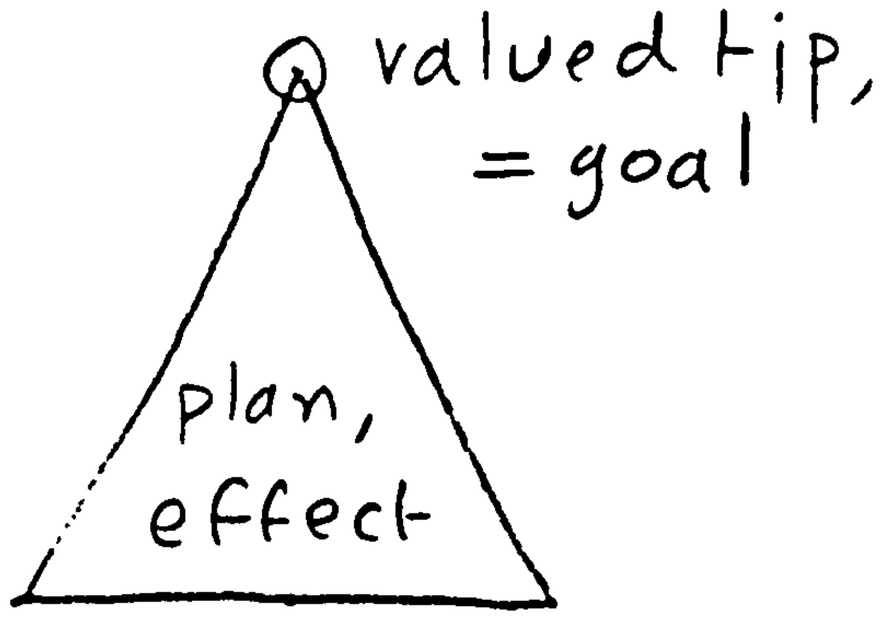
If an utterance's supporting or contradicting a precondition of a mutually known plan can affect the merit of the plan, what other relations between an utterance and a plan can be important? Can one decide between which sentences one should search for links? Is there any principle in it?

There is, and a simple one. It has to do with the ways of affecting the merit of a plan described in chapter 4. A plan has an effect if it makes sound (or unsound) an initially unsound (or sound) proof tree whose tip is either good or bad. The general relation between a plan and an effect is as in Search/1. This, and the following diagrams, are drawn in accordance with the appendix. The triangles are proof trees, and the overlap represents some node in the plan which is also a fringe node of the effect. The ordinary case, where the tip of the plan is valued and is the goal of the plan is the special case where the plan the effect wholly overlap and are

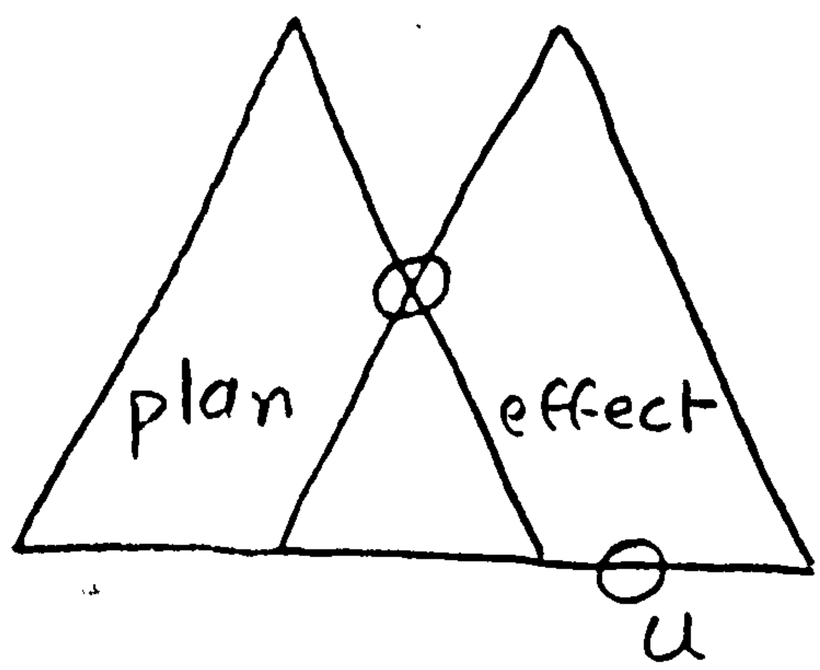
search/1



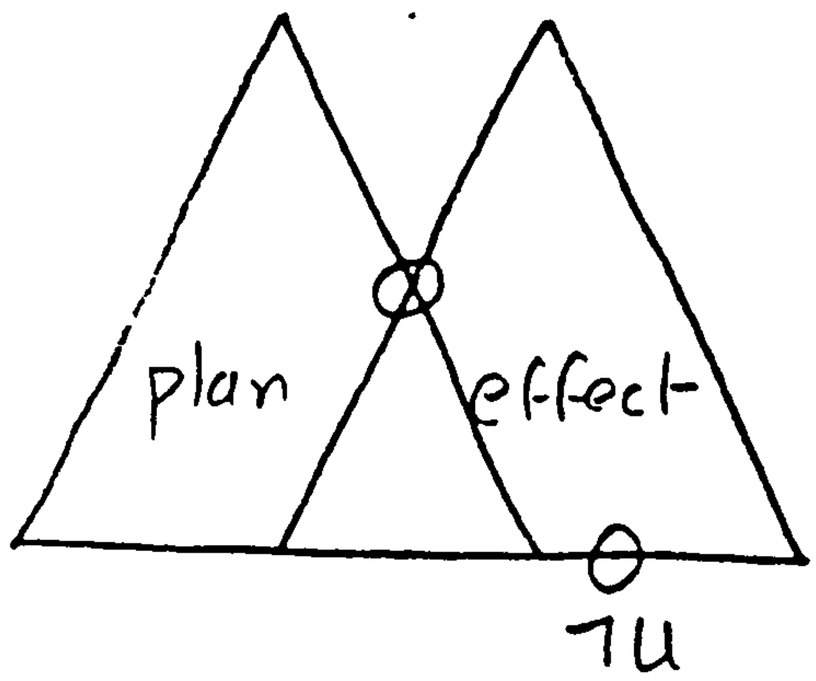
search/2



search/3



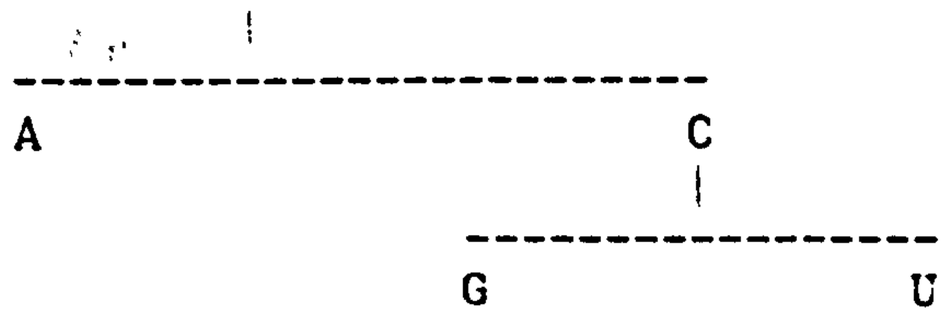
search/4



therefore identical, as in Search/2. For brevity I'll consider only the general case.

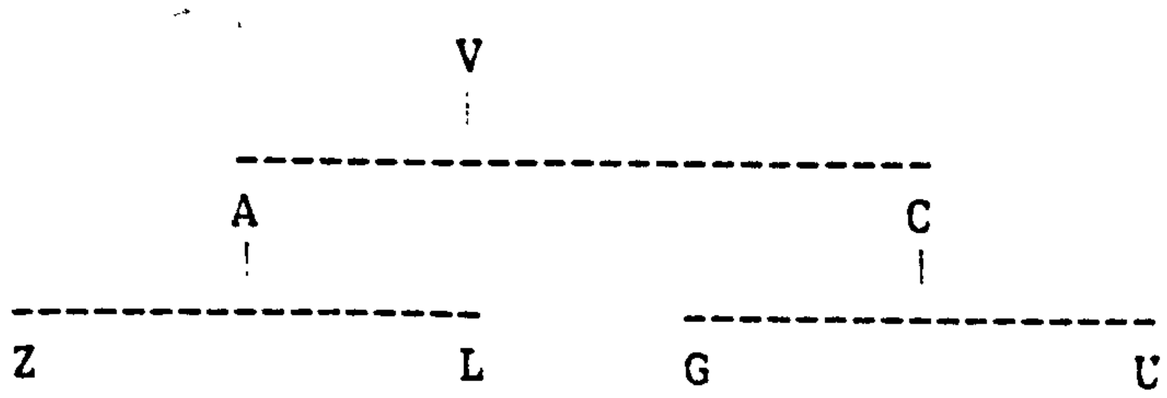
Given a statement of fact U and a plan, U may add an effect if all the preconditions of the proof tree that is the effect except U are already true, as in Search/3. Similarly it may remove an effect if all the preconditions in the effect are already true and if they include -U; as in Search/4; or more generally, if U entails the contradiction of a preconditon of the effect, as in Search/5.

This suggests what search must be done if one suspects that an effect is being added. From U one must search forward until one comes to a valued sentence that is not already proved, as in Search/6. The dashed line represents a series of inferences such as



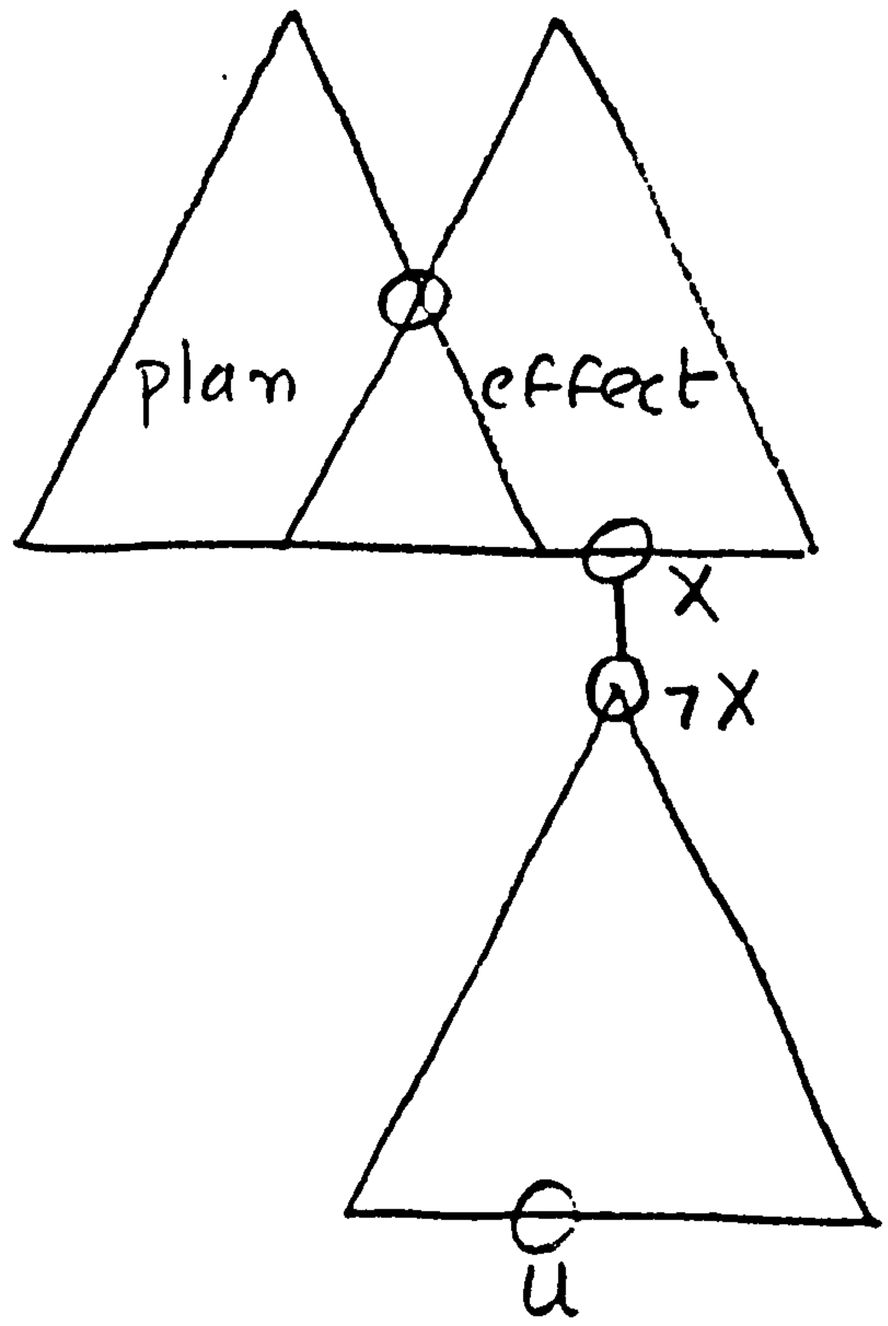
where the truth of A, G are at first irrelevant.

Then one tries to find a similar link between a node in the plan and the valued sentence; as in Search/7. Now the dashed lines may give you the skeleton of an effect proof tree: something such as

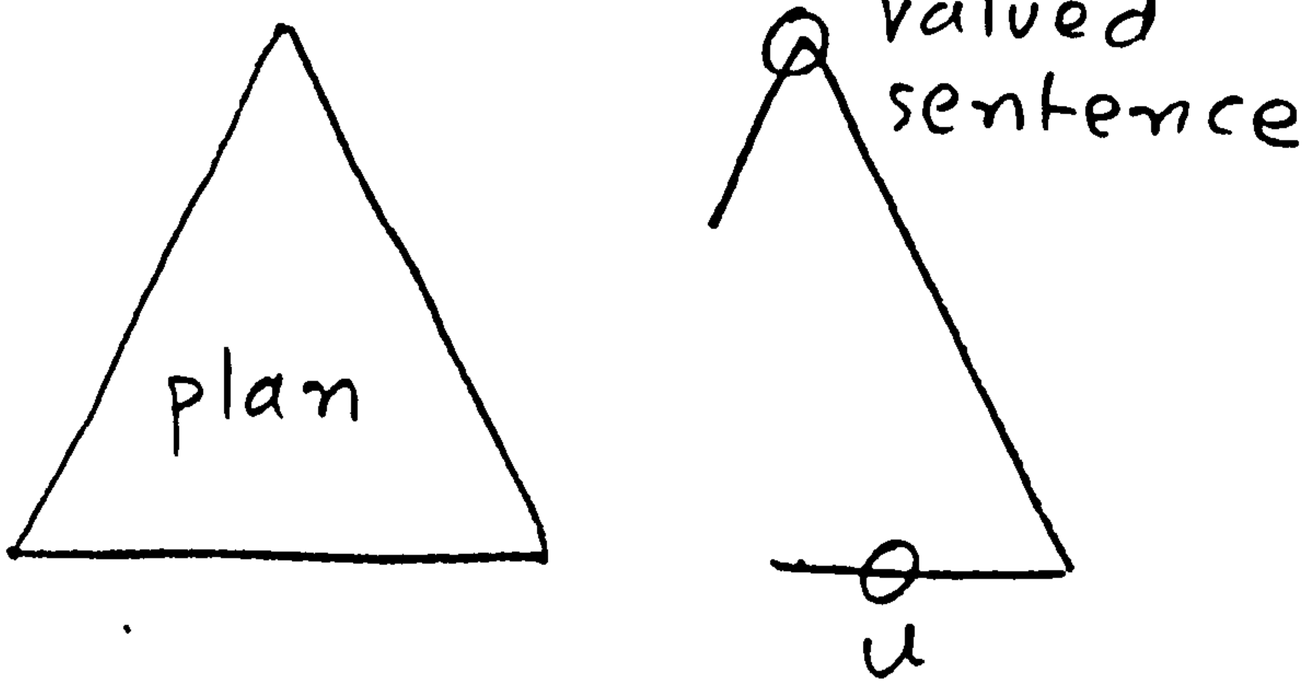


If one can hypothesize that Sp, the utterer of U believes L, G, and that he can believe that Hr can believe them too, then one can hypothesize that Sp intended was to add the new effect. Confirmation

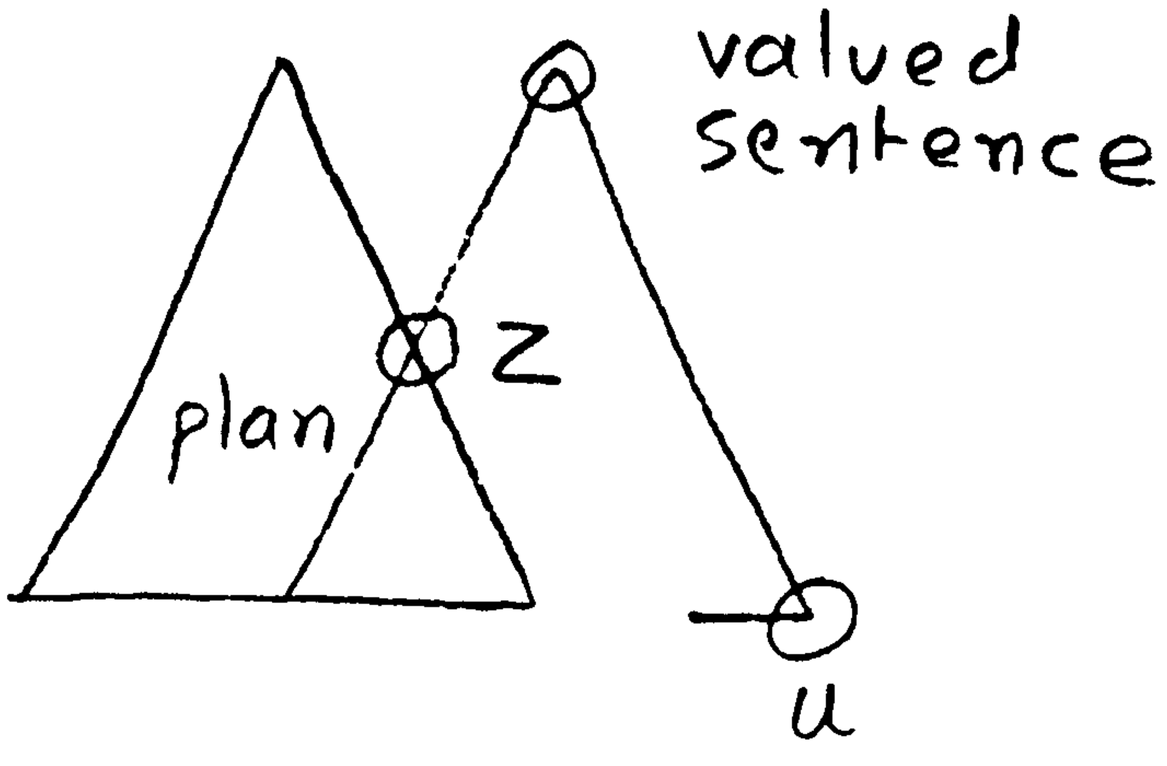
search/5



search/6



search/7



of this depends on being able to agree that adding the effect benefits Sp.

Looking for the removal of an effect is similar. One tries to find a link as in Search/8, or even as a two stage process as in Search/9, which is needed to cover an exchange such as

A: I'm going to the beach.

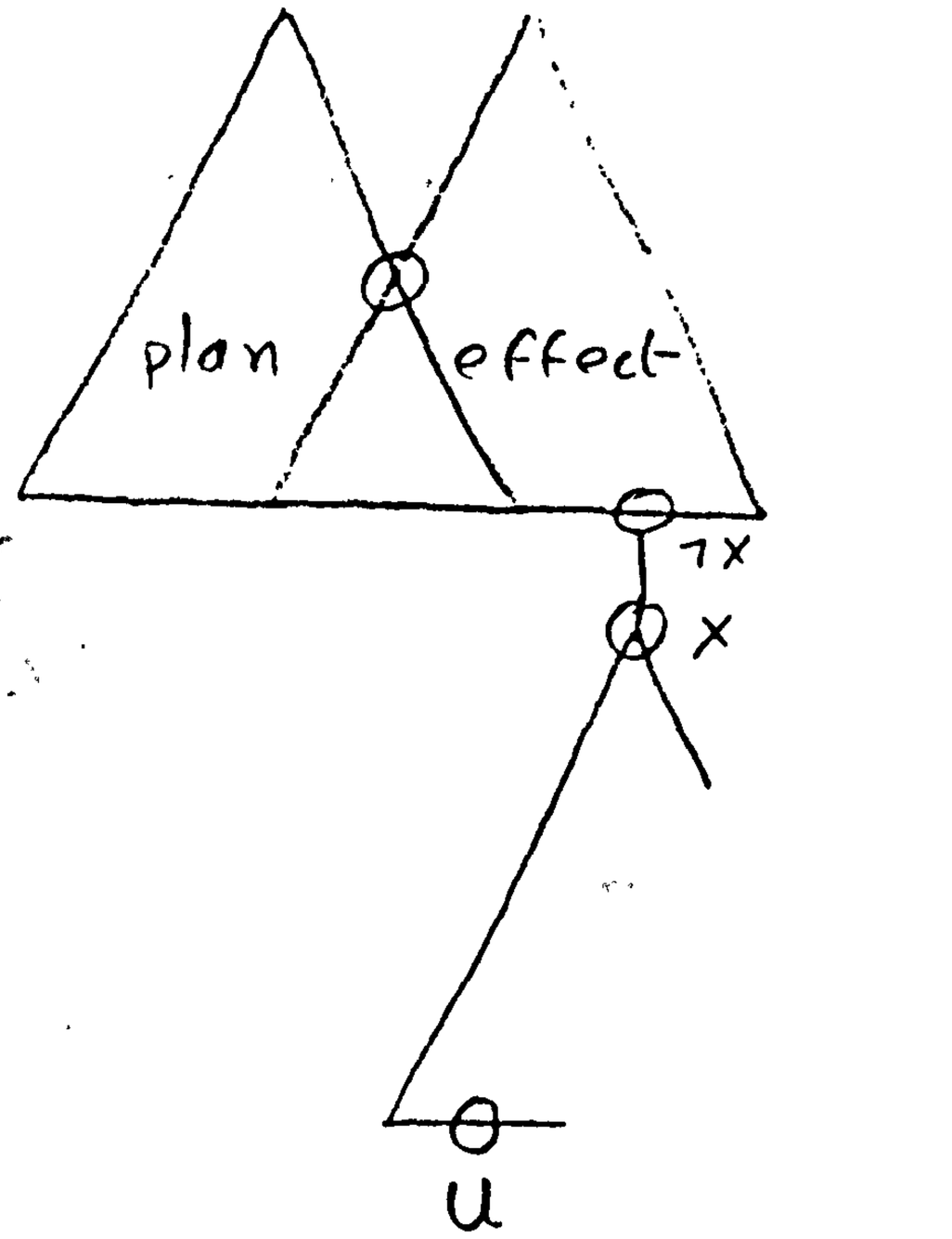
B: There's some sun tan oil in my bag.

That is an attempt to point out the possibility of removing a bad effect from a plan as suggested in Search/10.

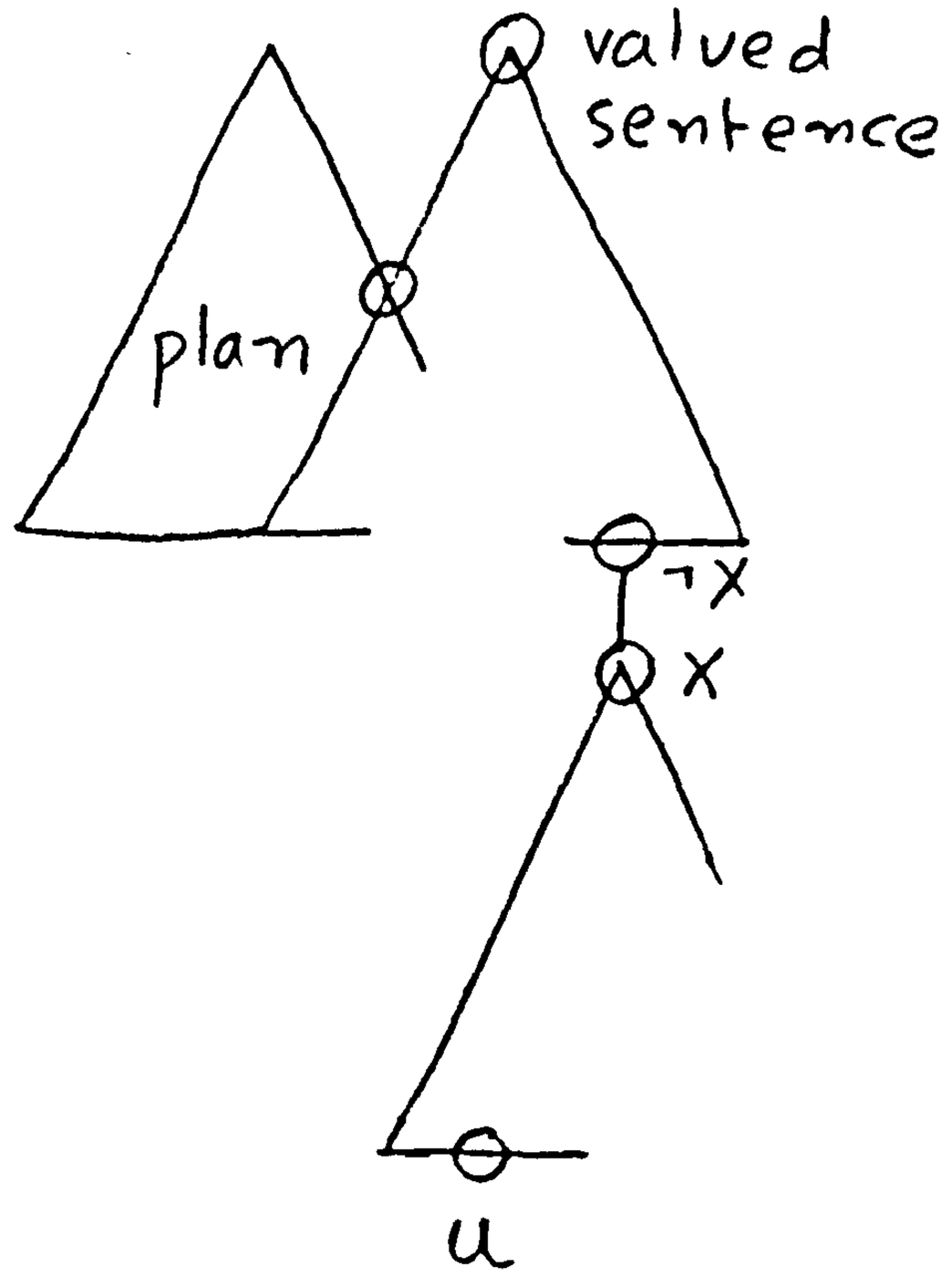
A statement of value has to be treated differently, but the principle is the same. A proof tree is only an effect if its tip is valued. Someone's statement of value can change what is known to be valued. Suppose Sp says "I want V". This will add an effect to a plan if one can find a link from the plan as in Search/11, and can accept the proof tree that the link suggests. Similarly "I don't want V" (taking it to mean "-(I want V)", not "I want -V") will remove the effect in Search/12.

Questions about fact or value, and commands, can be treated analogously if one assumes that it is the response to them that may affect a plan. For the question "Is U true?" one should try to link either U or -U to a plan in the way just described; both U and -U because either answer may be the important one. Similarly for the question "Do you want V?". One should try to see what both "I want V" and "I don't want V" do to the plan. For the command "Do F!" one should try to link both F and -F to the plan - again, either obedience or disobedience may be what matters.

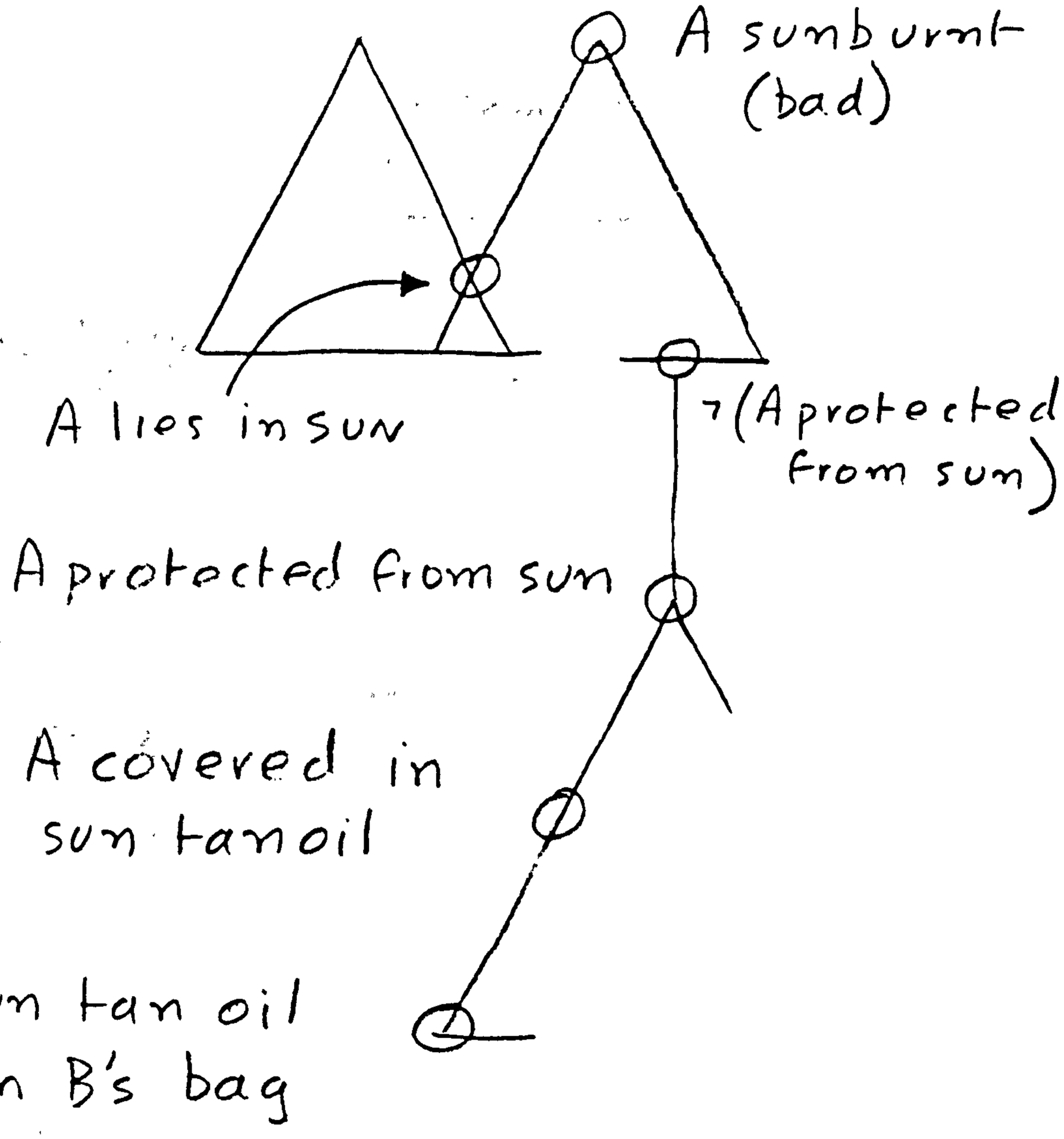
search/8



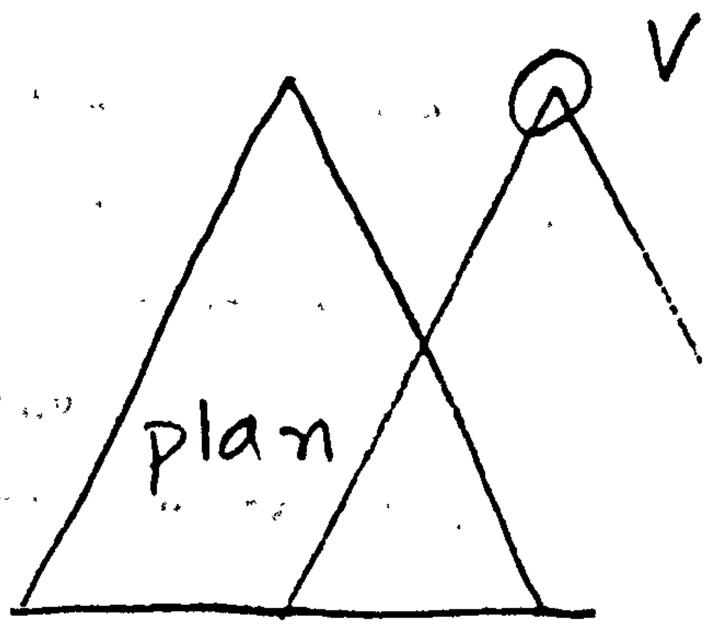
search/9



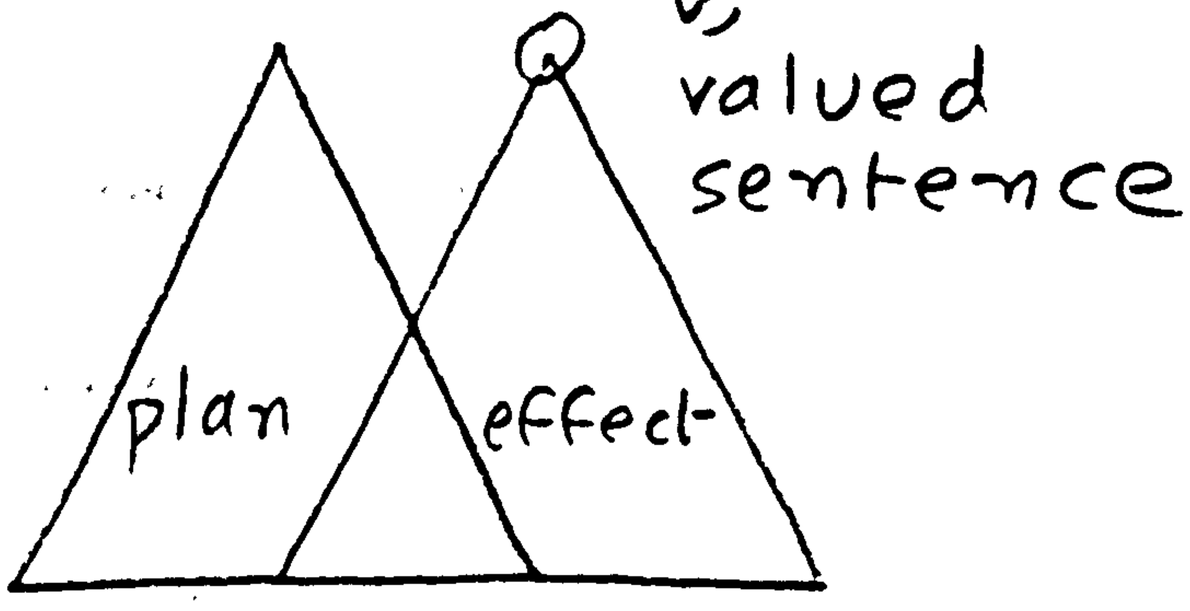
search/10


A sunburnt (bad)
A lies in sun
¬(A protected from sun)
A protected from sun
A covered in sun tan oil
sun tan oil in B's bag

search/11



search/12



search/13

he knows time

I know time

search/14

Sometimes, indeed quite often, an utterance will be made when there is no mutually known plan - as for instance when a stranger asks you "Do you know the time?". The essential thing is still to link the utterance via a link that may grow into a proof tree to a valued tip. Then if necessary one can hypothesize a plan. In the "Do you know the time?" example, one can hypothesize as in Search/13. This can be expanded very reasonably into one of the cases where the plan and the effect wholly overlap.

Sometimes though one has to guess the plan too, as in

A: Do you mind if I turn off the fire?

which can only be seen as a question relevant to the badness of an effect if one hypothesizes an action that might cause the effect, as in Search/14. The effect is the minimal proof tree of a single sentence.

The problem of course with this sort of branching search is controlling it. What I have said describes what is to be done (roughly, but formally it is too trivial to expand). But if the utterance U is an antecedent of many rules, the search will explode. Since my interest was trying to show how trivial processes could do interesting work, I always cheated, and tested examples where the rules that could be used in the search were more or less exactly those that I knew would be needed.

Nevertheless, this approach is probably the right one. It will, if it succeeds, find only real effects, and if it ran for ever it would find all effects. (There is an analogy with resolution theorem proving.) The heuristic that would probably be the most sensible to apply would be breadth-first searching. The utterance and all the

nodes in the plan would be the intial frontier of such a search. All sentences such that they were a conclusion of a rule of which some antecedent was in the current frontier would be the next frontier - and so on until the frontiers touched. Searches for contradiction can be made in an obviously similar way. Another heuristic is intuitively more appealing but more speculative. When the conversation turns on macintoshes, the goal of keeping dry is likely to motivate remarks. If one could make even a rough stab at proposing the valued sentences that might be the tips of effects, just on the grounds of the sort of things that had been mentioned, one would have three anchor points to attempt to chain between (plan, utterance, and also likely valued sentences), not just two. That should constrain search considerably.

A candidate method would be the demons of Charniak (1972, 1976). A demon is an inference rule which is only used in an attempt to make an inference once some of its antecedents have been already been established in the text being analysed. Other antecedents are then looked for explicitly. Not all such inference rules are available all the time. They must have been loaded by a "base routine", a process sensitive to the topics in a text. Which topics are known to be present depends on earlier inferences or explicit references.

The application to finding plan interactions would be that actions or states present in a plan tree would count as references to some topic. The topic's presence would lead to the loading of demon rules of the form "If person P's plan contains act A, then it is possible that F will occur". The detection of act A in a plan would fire the rule. Then a fact G would only be eligible for consideration as a valued tip in an interaction if "It is possible that G will occur" was currently true.

8.4. The status of theory and program

Allen's ideas are supported by the existence of a working program, whereas mine are not. Does this support one position against the other? There are two standard reasons for writing a program in AI research; the first is to demonstrate to other people that a superficially attractive theory does what it is claimed to do. The second is to help the researcher form his ideas to see where they are wrong.

Though there is no final program, there have been at least six programs of various degrees of wrongness that have been thrown away. They have been useful in two ways: in pointing out flaws, as above; and in making sure that case analyses were complete and that symmetries were caught.

For instance, at one stage I wrote code to look for the reasons that one might say "I want X", "I want -X" and "Do you want X?". This set of three cases suggested the hypothesis that they were three of the four cells of a two-by-two grid. The last cell would be "Do you want -X?". I couldn't think of any example of an utterance of this sort until I heard someone say "Do you mind if I turn the [electric] fire off?".

The question "Do you want X?" seemed to ask about the goal of an as yet unmade plan by the questioner for the answerer's benefit. The question "Do you mind if -X?" seemed to ask about the badness of some state about to be brought about by some made plan of the questioner's. But once these had been seen as symmetric cases, it seemed sensible to try and unify the code that would look for such connections. In each case this involved finding an action by the questioner which might have X or -X as a remote effect. To find such an action might involve ascribing to the questioner a plan that he

contemplated but had perhaps not yet resolved on.

It was already apparent that ascribing to a speaker a plan involved ascribing to him certain beliefs and values, and that doing this was the mechanism of indirect communication. The same code would extract such beliefs and values from that ascription of any plan. This fact suggested trying to see the effects of all utterances arising during a common stage of processing, where a plan would be ascribed to the speaker and then examined for what it entailed about his beliefs and desires. This idea was rejected, but it was not worthless. It was suggested by the desirablity of fusing blocks of code.

It would be silly to underplay the importance of having a working program, especially when, because of the difficulty of controlling search, many of the problems in AI are not "Can this inference be made?" but "Can this inference be made in practice?". Only working programs can demonstrate the latter. But even a program that never works can be a good way of finding out what it ought to do if it did.

Conclusion

To collect the conclusions of this thesis in detail would be to rehearse the introduction. But the major, the central, conclusion is that there are features of discourse which, if they are to be explained, require one to consider utterances as planned actions; and the sort of changes that those actions which are utterances bring about are changes in other peoples' plans.

There were four main arguments for this:

- There are general reasons why one might think that communication between independent agents could not employ utterances that were compulsory mind-changers, but would have to proceed by recognition of attempted plan alteration.

- Speakers such as ourselves produce intuitively coherent exchanges in which the coherence can be interpreted as the reasonable expectation that an utterance would produce beneficial changes in other peoples' plans, either for selfish or altruistic reasons. What the affected plan was could have been made clear by the previous discourse .

- There are discourse events which can be described by (for example) "is an acceptance"; "is a refusal", "is an explanation". It seems that the grounds on which these terms can be applied must refer to changes in the interlocutors' plans and knowledge about plans.

- Not all of the information given by from an utterance follows from its literal meaning: some of it can only be recovered if one looks at the change that the utterance compels in what must be assumed about

the speaker's plans. The speaker may employ this in indirect communication.

To support such a claim, one must propose in some detail a model of what plans are, how they are vulnerable to new facts and values, and what sorts of changes in people's plans are beneficial for the speaker. I gave a consciously simple but general account of plans which assimilated them to proofs. Changes in plans could then be achieved by the sort of changes that would make a proof valid or invalid. Besides this, plans involve a notion of the value of their effects, and I related effects to the value of the tip, and the validity of the body, of the proof tree that constitutes the effect.

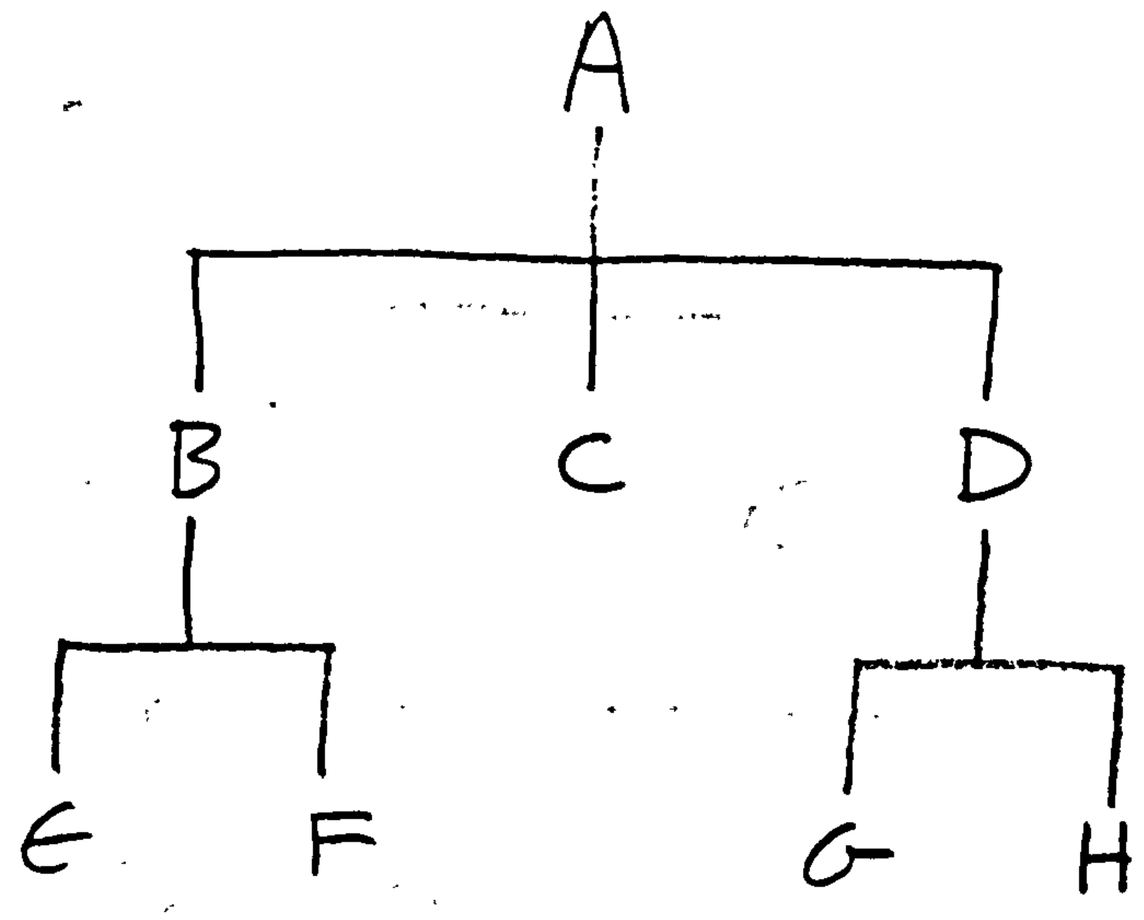
A second piece of detail needed is recursive belief. Any account of how people attempt to modify each others' plans must involve an account of how they represent each others' plans. I argued that in most cases this could be done in an unexpectedly simple way; people need pay attention only to the plans that they took to be mutually believed. Fully general recursive belief was not needed, since discourse typically fits what were demonstrated to be the conditions when mutual belief suffices.

Lastly, I sketched how such a theory could be made computational. Those sketches were not satisfactory, but that is a separate issue from the merits of the theory.

Appendix: Abbreviating proof trees

Much of this thesis is about how proof trees can interact. It will be convenient to have a graphic way of illustrating this.

A proof tree is roughly a tree in which each node is a proposition which is true if either it is at the fringe of the tree, or if all its daughter nodes are true. An entire proof tree such as Abbrev/1 can be abbreviated as any of Abbrev/2a-d. That is, the whole tree is drawn as a triangle. Particular nodes in the tree can be made explicit. If such an explicit node is a plan goal, it must be at the apex of a triangle; if it is a precondition, it must be at the base; if it is an intermediate state or act it must be on a side. A node that occurs in two trees can be drawn in both as long as it occurs in the right place in both. Thus the overlapping trees in Abbrev/3 can be abbreviated as Abbrev/4. If two nodes in different trees are the same except for opposite sign, that can be emphasized by a line drawn between them. So Abbrev/5 can be drawn as Abbrev/6. A tree, parts of which are not specified, so that the question of how some nodes are to be proved is left open, such as Abbrev/7, can be drawn as a partial triangle, such as Abbrev/8.



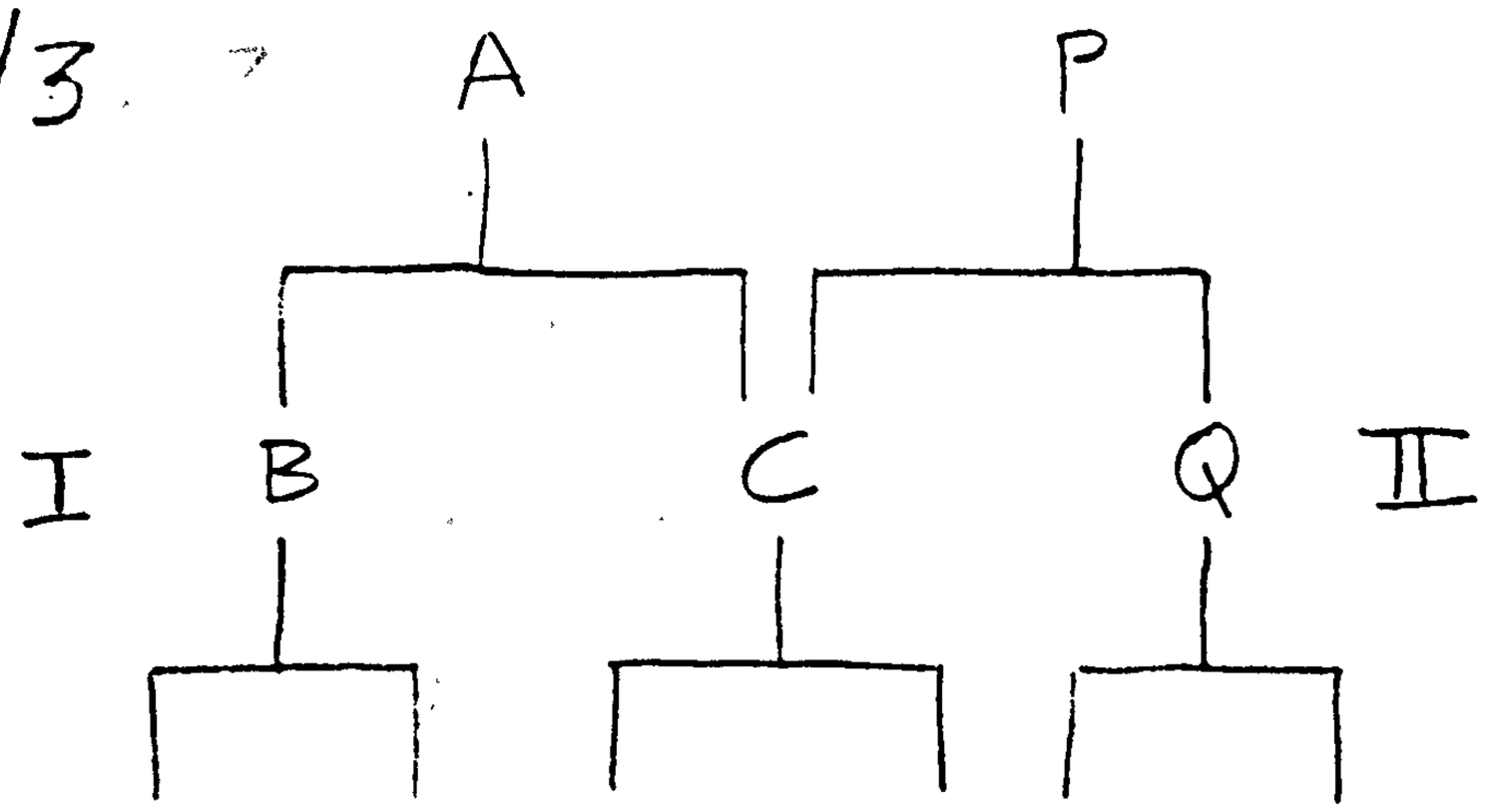
/2a

/2b

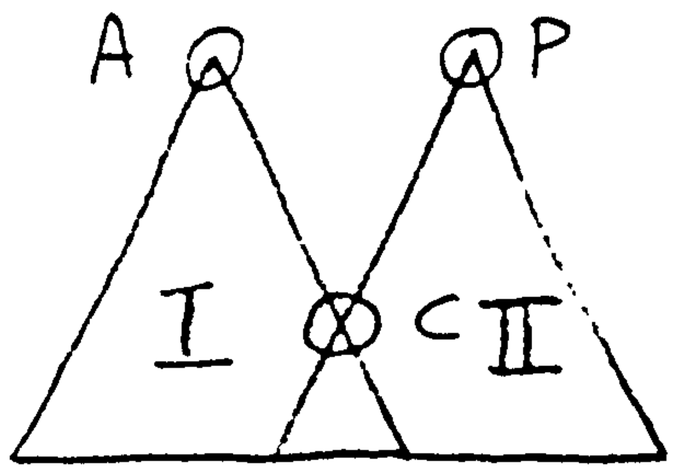
/2c

/2d

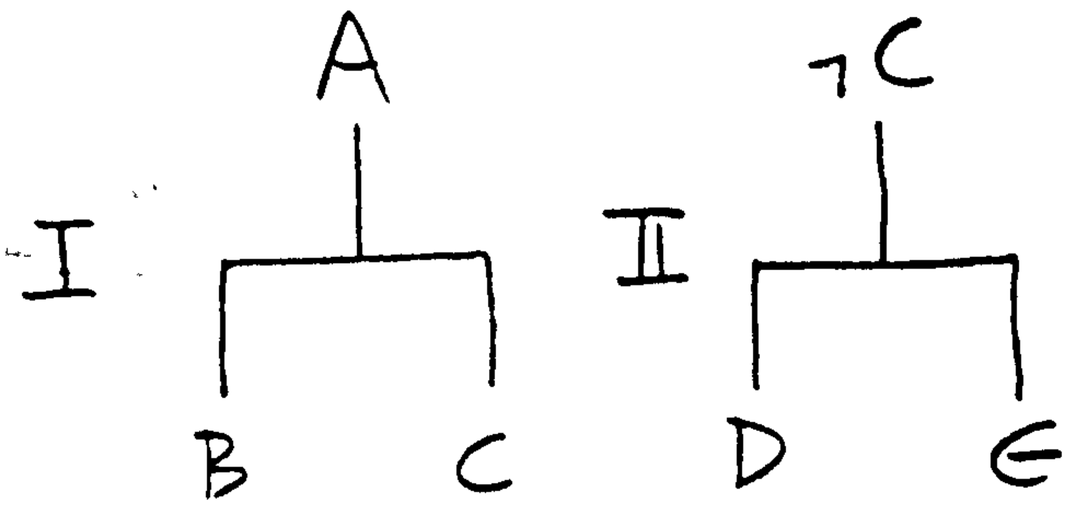
abbrev/3



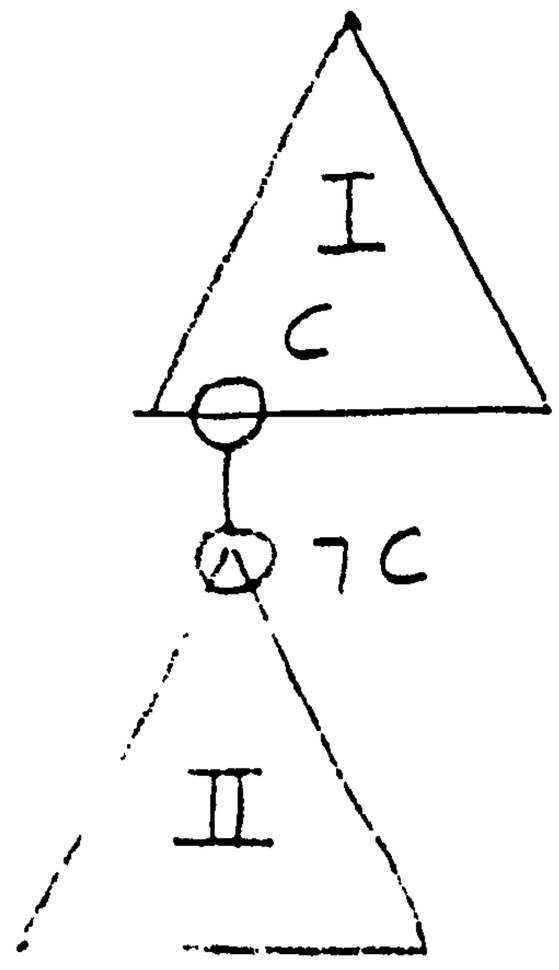
/4



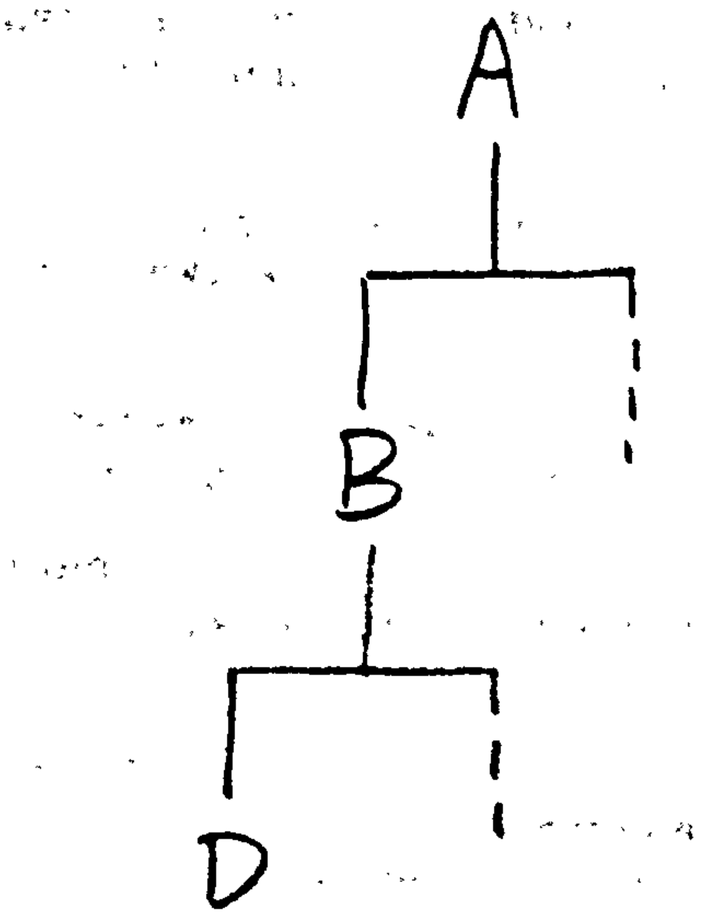
/5



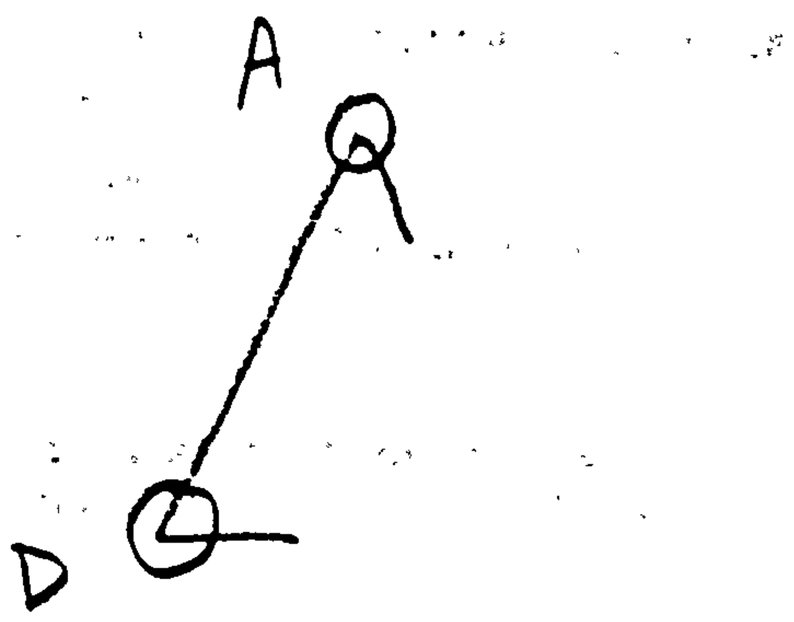
/6



abbrev/7



18



References

Allen JF:1979
A plan-based approach to speech act recognition
TR121/79, dept computer science, university of Toronto

Allen JF: 1983
Recognizing intentions from natural language utterances
in Brady & Berwick

Allen JF: 1982
ARGOT: a system overview, TR 101,
dept computer science, university of Rochester

Allen JF et al: 1983
ARGOT: the Rochester dialogue system,
in IJCAI-83

Allen JF, Perrault CR: 1980
Analyzing intention in utterances.
Artificial Intelligence, 15 (1980) 143-178

Austin JL: 1962
How to do things with words
OUP

Brachman RJ et al: 1979
Research into natural language understanding, annual report
BBN, report 4274

Brady M, Berwick R: 1983
Computational models of discourse
MIT

Brown GP: 1980
Characterizing indirect speech acts
J computational linguistics vol6#4

Charniak E, Wilks YA: 1976
Computational semantics
North-Holland publishing company

Charniak E: 1972
Towards a model of children's story comprehension
MIT AI Tech Report 266

Clark HH, Carlson T: 1982
Speech acts and hearers' beliefs.
in Smith ed 1982

Clark HH, Marshall CR: 1981
Definite reference and mutual knowledge.
In Joshi AK, Sag I, Webber B, eds

Cohen PR: 1978
On knowing waht to say: planning speech acts
TR 118, Dept computer science, university of Toronto

Cohen PR, Perrault CR: 1979
Elements of a plan-based theory of speech acts
in Cognitive Science: 3: 197-212.

Cohen PR, Levesque H: 1980

Speech acts and the recognition of shared plans
in CSCSI/SCEIO 1980

Cole P, Morgan JL, eds: 1975
Speech acts
Syntax and Semantics, vol 3
Academic press

Donellan K: 1966
Reference and definite description.
In Steinberg D, Jakobovits L, eds, Semantics, CUP.

Dyer MG: 1983
In-depth understanding
MIT press

Ellman J: 1983
An indirect approach to types of speech act
IJCAI-1983

Fish S: 1979
Normal circumstances, literal language,...
in Interpretive social science, eds Rabinow P, Sullivan WM
Campus

Gordon D, Lakoff G: 1975
Conversational postulates
in Cole & Morgan, eds, 1975

Grice HP 1957
Meaning
reprinted in Philosophical Logic, ed Strawson PF
OUP

Grice HP: 1968
Utterer's meaning, sentence meaning and word meaning
reprinted in Searle ed 1971

Grice HP: 1975
Logic and conversation
in Cole & Morgan eds 1975

Grosz B: 1977
The represenation and use of focus in dialogue understanding
Technical note 151, Artificial intelligence centre, SRI

Hobbs J: 1978
Why is discourse coherent?
SRI AI Centre Tech Note 176

Joshi AK, Sag I, Webber B: 1981
Elements of discourse understanding
CUP

Joshi, AK: 1982
Mutual belief in question-answer systems.
in Smith ed 1982

Lehnert W: 1982
Plot units: a narative summarization stategy
in Lehnet & Ringle, eds, 1982

Lehnert W, Ringle M, eds: 1982
Strategies for natural language processing.

Lawrence Erlbaum Associates.

Levinson, SC: 1983
Pragmatics
CUP

Lewis DK: 1969
Convention
Harvard university press

Perrault CR, Allen JF: 1979
A plan-based analysis of indirect speech acts
Report, Dept Computer Science, Toronto University
also in JACL 6(3) 167:182

Perrault CR, Cohen PR: 1981
It's for your own good: a note on inaccurate reference
in Joshi AK, Sag I, Webber B

Schank RC: 1980
Language and memory.
Cognitive science 4, 243-284

Schank RC: 1982
Remembering and memory organization.
in Lehnert W, Ringle M, eds 1982

Schank RC, Abelson RP: 1977
Scripts, plans, goals and understanding.
Lawrence Erlbaum Associates.

Schank RC, Riesbeck CK, eds: 1981
Inside computer understanding.
Lawrence Erlbaum Associates.

Searle J: 1969
Speech acts
OUP

Searle J ed: 1971
The philosophy of language
OUP

Searle J: 1975
Indirect speech acts
in Cole & Morgan 1975

Searle J: 1978
Literal Meaning
Erkenntnis 13 (1978)

Smith NV: 1982
Mutual knowledge.
Academic press.