SRI International

AN APPLICATION OF DEFAULT LOGIC TO SPEECH ACT THEORY

Technical Note 457

February 17, 1989

By: C. Raymond Perrault, Director
Artificial Intelligence Center
Computer and Information Sciences Division
and
Center for the Study of Language and Information
Stanford University

APPROVED FOR PUBLIC RELEASE:
DISTRIBUTION UNLIMITED

This work was supported in part by a gift from the Systems Development Foundation.

SRI International®

333 Ravenswood Ave. • Menlo Park, CA 94025
(415) 326-6200 • TWX: 910-373-2046 • Telex: 334-486

Revised version of CSLI Report 87-90 (March 1987)

To appear in P. R. Cohen, M. Pollack and J. Morgan, (eds.), Intentions in Communication, MIT Press.

An Application of Default Logic to Speech Act Theory

C. Raymond Perrault

Artificial Intelligence Center and
Center for the Study of Language and Information
SRI International

Abstract

One of the central issues to be addressed in basing a theory of speech acts on independently motivated accounts of propositional attitudes (belief, knowledge, intentions, etc.) and action is the specification of the effects of communicative acts. The very fact that speech acts are conventional means that specifying the effects of the utterance of, say, a declarative sentence, or the performance of an assertion, requires taking into consideration many possible deviations from the conventional use of sentences – specifically uses that are insincere, not serious, or indirect. Previous approaches to the problem of specifying speech act consequences have paid insufficient attention to the dependence of the participants' mental state after an utterance on their mental state preceding it. We present a limited solution to the problem of belief revision within Reiter's nonmonotonic Default Logic and show how to formulate the consequences of many uses of declarative sentences. Default rules are used to embody a simple theory of belief adoption, action observation, and the relation between the form of a sentence and the attitudes it is used to convey.

Introduction

Speech act theory aims to relate three aspects of utterances:

- The type of actions they are being used to perform, such as asserting and convincing.

- The syntactic and semantic features of the utterances, such as their sentence type (declarative, interrogative, imperative), propositional content, and intonational pattern.

- The state of the world before and after the utterance, particularly the mental state of participants and observers.

After arguing that all utterances should be viewed as actions of the speaker, Austin [1962] proposes three levels of speech act description. In saying "It's cold here" I utter a well-formed sentence of English, with definite sense and reference. Austin calls

this act *of* saying something the *locutionary* act. In so saying, I may also be asserting that it is cold, or even asking you indirectly to turn the heat up. These actions performed *in* saying something Austin labels *illocutionary* acts. Finally, I may convince you that it is cold, or even that you should turn up the heat. These actions performed *by* saying something he calls *perlocutionary* acts. A speaker can perform an illocutionary act successfully while failing to carry out a related perlocutionary act. Thus, I can assert successfully that it is cold here without convincing you of that fact. One important feature of illocutionary verbs is that they can be used in so-called explicit performative sentences such as "I hereby *assert* that it is cold here," whereas perlocutionary verbs cannot.

In some of the most widely read literature on the subject (e.g., [Searle 1969]), speech act theory seems, at first glance, reducible to the problem of characterizing the successful performance of illocutionary acts, i.e., of specifying conditions under which the performance of an action, and, in particular, the utterance of one or more sentences can be interpreted as the successful performance of a given speech act. Finding a solution to this problem is a useful contribution to the lexical semantics of the language, but is interesting mostly as a route to a general understanding of the difference between actions that are inherently communicative (illocutionary acts) and those that can be achieved by noncommunicative means (perlocutionary acts). Iago, for example, convinces Othello of Desdemona's infidelity by leaving her handkerchief on his path. Clearly, he does not thereby *communicate* anything to the Moor.

Illocutionary acts can be performed *mistakenly* (as in assertions whose content happens to be false), *insincerely* (as with lies), *indirectly* (as in using "It's cold here" as a request to turn on the heat) and *nonseriously* (as in using "This was a terrific meal" to mean that it was terrible.) Largely because of this range of uses, it is difficult to give direct definitions relating the form of utterances to the illocutionary acts they are used to perform. Searle's account says little on the subject, save his claim that the hearer's recognition of the speaker's intentions is to be "in virtue of (by means of) [the hearer's] knowledge of the meaning of [the utterance]." He discusses indirect speech acts further in Searle [1975].

Searle views illocutionary acts primarily as moves made as part of a larger social activity while their felicitous performance he regards as being governed by what he calls constitutive rules. While this approach may be necessary for a full account of "institutional speech acts," such as betting, convicting, and marrying, it has not been very useful in explicating the relation between the form and function of utterances. To address that issue, Allen, Cohen and Levesque, and I have been developing versions of speech act theory that place the burden on a direct account of the effects of utterances on the mental state of the participants. Illocutionary acts are relegated to a secondary role.

Cohen and Levesque [1985] (hereafter C&L), propose a theory of speech acts in four parts:

- An account of some propositional attitudes (e.g., belief, knowledge, intention),
- A theory of action and its relation to the attitudes, describing those necessary to engage in action and those resulting from it,
- A description of the effects of locutionary acts on the mental state of the participants, i.e., of sentences with particular syntactic and semantic features,
- Definitions of the performance of illocutionary acts as the performance of *any* action, linguistic or otherwise, under appropriate circumstances, by a speaker holding certain intentions.

C&L's account, couched in a logic of attitudes and action, is by far the most detailed and precise proposal yet. They provide a formal language in which to axiomatically state properties of some propositional attitudes, actions, and time, and they give the language a formal semantics. However, it still fails in two crucial respects: it makes arguably wrong predictions in some circumstances and predictions that are too weak in others. Although we focus on C&L because of its detail, earlier proposals, such as those of Cohen and Perrault [1980] and Perrault and Allen [1980] suffer from the same flaws. The heart of the problem is that the mental state of the speaker and hearer after an utterance is strongly dependent on their mental state before. Several examples are given in the next section.

These problems, I claim, can be overcome by defining many of the consequences of speech acts as *defaults* that are assumed to hold as long as there is no evidence to the contrary. I will show that a more perspicuous account of speech act consequences can be derived from a default theory describing

- the change in an agent's attitudes of over time,
- the transfer of attitudes between agents,
- the consequences of an action on an observer's mental state,
- the assumption that actions are intentional,
- the assumption that the utterance of sentences with particular features reveals particular aspects of the speaker's mental state.

In the next section we expand upon the problems encountered in specifying speech act consequences. We then postulate a simple theory of the change of beliefs over time, which we use to discuss informally what should be the result of performing declarative sentences in a variety of contexts. These are the facts our theory will explain. We then introduce the requisite notions from default logic, present the default theory necessary for speech acts and explore some of its consequences. Finally we examine a number of methods of defining illocutionary acts in this framework and outline systematic differences among various uses of utterances that, in different ways, do not conform to simple correspondences between form and function.

The Importance of Attitude Revision

A theory of speech acts should at least account for the fact that communicative acts must be overt, and that moods can be used in non-standard ways, as suggested in the preceding section.

Before we present C&L's axioms for imperative and declarative sentences, some notation will be helpful. If p is a proposition and x an agent, we take $K_x p$, $B_x p$ and $G_x p$ to mean that x *knows* that p, x *believes* that p, and x has the *goal* that p. We say that x and y mutually know that p, notated $MK_{x,y}p$, if and only if $K_x p$ & $K_y p$ & $K_x K_y p$ & $K_y K_x p$ & ... Mutual belief is defined analogously. A one-sided version of mutual belief is also useful: $BMB_{x,y}p$ if and only if $B_x p$ & $B_x B_y p$ & $B_x B_y B_x p$ & ... , i.e., if x believes that p and that x and y mutually believe that p. It will be convenient to think of mutual knowledge and belief formulas as the set of their conjuncts.

In monotonic logics, the consequences of actions are specified by axioms of the form "If p, then, after action α, q". We will refer to p as the *gating condition* and q as the *consequent*. Exceptional conditions are handled by stating axioms with different gating conditions and by specifying the strongest condition that holds after all instances of the action that are performed in states satisfying the gating conditions.

C&L's only axiom for the use of declarative sentences can be glossed as follows: if it is mutually known by speaker S and hearer H that e is an event of utterance of the sentence s to H, that S is the agent of e, and that s is a declarative sentence with propositional content p, then, after the utterance, it becomes the case that the hearer believes that it is mutually believed (BMB) that the speaker intends the hearer to recognize his intention that the hearer believe that the speaker believes that the propositional content of the declarative is true.

$$MK_{S,H}(\text{Utter}(H, s, e) \,\&\, \text{AGT}(S, e) \,\&\, \text{Attend}(H, S) \,\&\, \text{declarative}(s, p) \supset \text{after}(e, BMB_{H,S} G_S B_H G_S B_H B_S p))$$

They give a similar axiom for imperative sentences.

$$MK_{S,H}(Utter(H, s, e) \,\&\, AGT(S, e) \,\&\, Attend(H, S) \,\&\, imperative(s, p) \supset after(e, BMB_{H,S}G_SB_HG_SB_HG_SG_Hp))$$

In the declarative axiom, Utter(H, s, e) & AGT(S, e) & Attend(H, S) & declarative(s, p) is used as a gating condition. In a full theory of action and attitudes, these axioms must, along with others, allow the inference, under "normal" conditions, that B_Hp is a consequence of a declarative sentence with propositional content p, or that G_Hp is a consequence of an imperative. This second set of axioms must then, under some circumstances, allow for B_Hp to follow from

$$BMB_{H,S}G_SB_HG_SB_HB_Sp$$

When the speaker is mistaken or lying, B_Hp is still a consequence, as long as the hearer does not recognize the mistake or detect the lie. In the first case, he might come to believe that B_Sp, that $G_SB_HB_Sp$ and that G_SB_Hp. In the second, he should come to believe that $\neg B_Sp$, that $G_SB_HB_Sp$ and that G_SB_Hp. It is unnecessary to delve here into what the axioms need be.

Unfortunately, the consequent of the declarative axiom is still too strong to hold in all conditions of utterance. Consider, for example, the ironic use of "This is the best meal I ever had." It is easy to verify that none of the following conditions hold after the utterance (where p, of course, is the propositional content of the utterance.)

B_Sp
B_HB_Sp
$G_SB_HB_Sp$
$B_HG_SB_HB_Sp$
$G_SB_HG_SB_HB_Sp$
$BMB_{H,S}G_SB_HG_SB_HB_Sp$

Thus, in the nonserious cases, the predictions of the declarative axioms are too strong.

This observation might suggest that some modification of the consequents of the axioms might suffice, but such is not the case. The main difficulty arises from the fact that the consequents of the declarative and imperative axioms are unaffected by the mental state of speaker and hearer *before* the utterance: the theory only places constraints on their mental state afterwards.

Consider what one might take as the simplest possible context-independent consequence of a declarative sentence, namely, that H believes p. Now, of course, this is unacceptable, as H might not be convinced by the utterance: H might have good reasons, for example, to continue to believe not p. In the context-insensitive accounts, this problem might be handled by having the more complex attitude $B_H B_S p$ as the innermost subformula of the consequent. But this move merely defers the problem: H might previously have had reason to not believe that S believes p, as would be the case, for example, if H had just seen S observe a physical situation in which not p was true. By a similar argument, no formula of the form $B_H B_S B_H ... p$ can be taken to hold after *all* uses of declarative sentence, since imposing $B_S B_H ... p$ as a belief of H after the utterance may conflict with the beliefs he had before.

The next possible way out is to weaken the consequent by having the hearer comes to believe that the speaker has certain goals or intentions, not just beliefs. In the simplest case involving declaratives, this would be having H come to believe that S has the goal that H believe p. i.e., $B_H G_S B_H p$. Again there is trouble ahead, as one would suspect from the fact that one is still postulating a *belief* of the hearer's. The only way this could work is if it were possible for H, after the utterance, to acquire a belief about S's goals that could never contradict H's mental state before the utterance. However, this hope vanishes once it is recognized that one must assume that an agent cannot hold goals (or intentions) that he believes to be impossible. Thus, if the speaker believes before the utterance that the hearer believes not p, and would not change his belief even if the speaker asserted p, then the speaker cannot have the goal that the hearer should come to believe p as a consequence of S's uttering a declarative sentence with content p. Therefore, for H to come to believe, after the utterance of the declarative p, that S's utterance reflected his goal that H believe p is inconsistent with H's believing (both before and after the utterance) that S believes that H believes not p. Thus H's recognition that S intends H to recognize that S believes p cannot be true after *all* utterances by S of a declarative p. This argument can be extended to show that *no* formula of the form $B_H G_S q$ can be taken as a context-independent consequence of the utterance of a declarative sentence.

One must therefore conclude that the consequences of declarative sentences must depend on the mental state of the participants before the utterance. It is possible that one could formulate context-dependent axioms, as Cohen and Levesque attempt in Cohen and Levesque [this volume]. We follow the route of non-monotonic systems as it is in some ways simpler, cleaner, and more revealing of the relation between utterances and attitude revision. A brief discussion of some of the differences between the approach presented here and that of Cohen and Levesque [this volume] can be found near the end of this paper.

The Persistence Theory of Belief

Speech acts reveal certain aspects of the speaker's mental state, and cause changes in the state of the hearer(s) that are based on their perception of the state of the speaker. An agent's beliefs after an utterance, for example, will, in general, depend on his beliefs before it, as well as on its content. Ideally, one would like to have a theory in which it is possible for one agent's beliefs, say, to change according to how strongly he believed something before the utterance, as well as on how much he believes what the speaker says. We cannot give such an account in detail, so we shall rely on something simpler. We assume what might be called a *persistence theory of belief*: that old beliefs persist and that new ones are adopted as a result of observing external facts, provided that they do not conflict with old ones. In particular, we assume that an agent will adopt the beliefs he believes another agent has, as long as those don't contradict his existing beliefs.

Let us examine the consequences of this assumption for speech acts. We shall do so from the perspective of an observer who initially has a [partial] theory of the mental states of both speaker S and hearer H. What is to be accounted for is the observer's picture of S's and H's mental state after the utterance. We limit ourselves at first to a consideration of the participants' beliefs; intentions will be discussed later.

Example 1. [Sincere assertion] Let a speaker S address a declarative sentence with propositional content p to a hearer H in an initial state in which S believes that p, and in which neither S nor H has other beliefs about p or about the beliefs others hold regarding p. In declaring that p, S "expresses" his belief that p. We would like to claim here that, after the utterance, S still believes p, and as that, as a consequence of the utterance, H comes to believe that S believes p; thus H comes to believe p as well. If we also assume that S believes that H was observing S's utterance, then as long as S has no reason to believe that H does not believe that S believes p, S will also come to believe that H believes that S believes p; thus S will come to believe that H believes p. Continuing this argument, it should be true after the utterance that S believes p, that S believes that H believes p, that S believes that H believes that S believes p, and so on, and that H believes p, that H believes that S believes p, that H believes that S believes that H believes p, and so on again. Taken together, these progressively more complex formulas describe what has been called *mutual belief* by S and H that p. Mutual belief and its analogue for knowledge are discussed at length by Schiffer [1972] and Smith [1982].

Two points should be noted here. First, mutual belief occurs only by virtue of the assumption that H was observing S's actions, and that S was observing H observing S, and so on – namely, solely because of the overtness of the utterance. Thus, in the simpler situation in which, say, S is lifting a rock and H is observing S without being

observed. in turn, by S. S comes to believe that the rock has been lifted, H comes to believe that S believes that the rock has been lifted, but S has no further beliefs about H. Second, we are not claiming that the justification we gave for why it was rational for each agent to have certain beliefs had to be computed in the same step-by-step fashion by the agents themselves. One can think of overt actions producing mutual belief directly, in virtue of mutual observation of the situation. We are simply trying to describe the joint mental state of S and H, ignoring for now the structures necessary to implement this mental state and the procedures necessary to change them.

Now let us look at a slightly more complex example.

Example 2. [Lie] Suppose now that S utters the same declarative sentence, but this time believing $\neg p$; suppose also, as in Example 1, that S and H have no further beliefs about p. From the assumption of persistence, we would expect S to continue to believe $\neg p$, as his beliefs about p should not be affected by what he says about it. However, as before, we would expect H to come to believe that S believes p and therefore H to come to believe p. S would have no reason not to believe that H believes that S believes p, and thus no reason not to believe that H believes p. Thus in the state following the utterance, S and H mutually believe that p, except that S believes $\neg p$ instead of believing p.

Example 2 is just one of many examples of ways in which the context of utterance affects the resulting state. Here is a slightly more complex case.

Example 3. [Unsuccessful lie] Consider the following sequence of events. Felix is a very fat cat who can jump down from chairs but never climbs up anything. Suppose H enters a room behind S, who does not notice him. H sees S put Felix on a chair. S looks away from Felix (still not noticing H) whereupon Felix jumps to the floor, noticed by H but not S. H then leaves the room, still unobserved. S now leaves the room and meeting H in the hall tells him "Felix is on the floor." Let p be the proposition "Felix is on the floor." Prior to the utterance, S believes $\neg p$, H believes that S believes $\neg p$, but H believes p. After S's utterance, S continues to believe $\neg p$, while H continues to believe p and that S believes $\neg p$. However, S comes to believe that H believes that S believes p, so that S also comes to believe that H comes to believe p. H has no reason not to believe that S believes that H believes that S believes p, and also comes to believe that S believes H believes p. The final state is that S and H mutually believe p, except that S believes $\neg p$, and H believes that S believes $\neg p$.

Example 4. [Irony] In our final example, we assume that it is initially mutually believed that $\neg p$. S's declaring that p, as in saying "This was a wonderful meal" after one that was patently not, has no effect on the beliefs held by the participants have

about the quality of the meal. Thus the state after the utterance is that S and H still mutually believe $\neg$p.

Default Logic

The monotonic accounts of speech acts start with axioms that describe complex consequences of utterances – these consequences to hold in all contexts satisfying some condition. The consequences are complex, in the sense that they consist of deeply nested formulas and that the "simple" consequences (e.g., that the hearer believe the content of the assertion) must be inferred from further conditions (e.g., that the speaker is [believed to be] sincere.) The account we propose turns the formula complexity picture on its head: it assumes that the utterance of, say, a declarative sentence, indicates that the speaker believes its content p, but that this consequence occurs *by default*, only if it is not contradicted by the context of the utterance. Similarly, the fact that the hearer comes to believe p may follow his believing that the speaker believes p, as long as the hearer's believing p is consistent with his other beliefs.

The formal account follows Reiter's Default Logic [Reiter 1980]. Default Logic was developed to deal with the following kind of reasoning.

(1) Birds (normally) fly
(2) Tweety is a bird
(3) Penguins don't fly
(4) Penguins are birds
(5) Tweety is a penguin

From (1) and (2), it should be possible to *assume*, by default, that Tweety flies, as no available information prevents the assumption from being made consistently. Adding (3) and (4), it should still be possible to assume that Tweety flies, as there is no evidence that Tweety is a penguin. From (1)-(5) however, it should follow that Tweety doesn't fly. The reasoning demonstrated here is *nonmonotonic*: adding (5) to (1)-(4) cancels the conclusion that Tweety flies. Note also that, if (1) and (3) are jointly represented as

$$\forall x\ \text{bird}(x) \,\&\, \neg\,\text{penguin}(x) \supset \text{fly}(x),$$

then fly(Tweety) is not a *logical consequence* of (1)-(4), athough it is a default consequence.

Reiter's approach to default rules is to take them as rules of inference rather than axioms. Given a base language L (for the moment first-order predicate calculus), a *default rule* is of the form

$$\frac{\alpha : M\beta}{\omega}$$

where α is the *prerequisite* of the rule and w its *consequent.* The rule is read: if α is believed and β is consistent with what is believed, then w can be assumed. Of particular importance are rules of the form

$$\frac{\alpha : M\beta}{\beta}$$

which are called *normal defaults.* All the default rules used in this paper are normal; we shall abbreviate the rule above as $\alpha \Rightarrow \beta$. A *default theory* $\Delta = (D,W)$ consists of a set D of defaults and a set W of well-formed formulas of L called the *assumptions.*

In part because it is possible to have both $\alpha \Rightarrow \beta$ and $\alpha \Rightarrow \neg\beta$ as default rules of the same default theory, the notion of consequence in Default Logic is different from the notion of theorem in monotonic logical theories. The appropriate notion here is that of an *extension.* Given a default theory Δ, an *extension* of Δ is a set of formulas E which is the closure of W under the defaults D. Reiter suggests that E should have the following properties:

- It contains all the assumptions, namely, $W \subseteq E$.

- It is closed under the logical consequence relation of the base language, which we write $Th(E) = E$,

- It is closed under the default rules, in the sense that it contains the consequents of defaults whose antecedents it also contains, as long as the extension is consistent with the consequents, where "consistent with" means "does not also contain its negation": If $\alpha \Rightarrow \beta$ is in D, α is in E, and $\neg\beta$ is not in E, then β is in E.

Reiter's definition of an extension is given in terms of a fixed-point operator. However, he gives an alternative characterization that is better suited to our purposes.

DEFINITION. Let $E \subseteq L$ and $\Delta = (D,W)$. Let $E_0 = W$, and for all $i >= 0$,

$$E_{i+1} = Th(E_i) \cup \{\beta \parallel \alpha \Rightarrow \beta \in D, \text{where } \alpha \in E_i, \text{ and } \neg\beta \neg \in E\}.$$

Then E is an extension for Δ iff $E = \cup E_i$, over all i. $\oplus$

Because of the occurrence of E in the definition of E_{i+1}, the definition of extension is nonconstructive: it gives a rule for *verifying* that a theory E is an extension of a default theory Δ. This limitation is, unfortunately, the price one pays for the sort of nonlocal reasoning that nonmonotonic logics allow. In contrast to the unique set of theorems of monotonic theories, default theories may have several extensions, each of which is a consistent way of applying the defaults to the assumptions. Reiter shows that theories containing only normal defaults can be shown to have at least one extension; if they have more than one, these must be mutually inconsistent.

Default Rules for Speech Acts

We can now begin to develop our account of the role of default reasoning in speech act theory, restricting ourselves here to declarative sentences. We shall present a set of default rules D that will be applied as follows. Let W be a theory describing an observer's knowledge of the state of the world before an utterance is performed. W shall contain propositions labelled with the time, say 0, at which the utterance is performed. Let W' be W augmented by the statement that an utterance has been performed at time 0 as well as by statements describing the facts about who is observing whom at that time. We claim that each extension E of the default theory (W', D) describes a consistent view of the observer's knowledge of the world both before and after the utterance. Generally, the extensions will contain additional information about the beliefs of the participants after, and about the intentions of the speaker before the utterance.

First we define a language L to express facts about agents, actions, and propositional attitudes. L has expressions of type *agent*, *time*, *action*, *untensed proposition* and *tensed proposition* (which we call simply *proposition*). The constants S and H and the variables x and y are the only expressions of type *agent*. The integer constants 0, 1, ... and the variables t and u are of type *time*. If p is of type *proposition*, then p. is of type *action* and denotes the action of uttering a declarative sentence with [tensed] propositional content p. If z is of type *agent*, Obs(z) is of type *action* and denotes the action of observing z.

The expressions of type *untensed proposition* consist of propositional variables r, s, ... and their boolean combinations. The expressions r_t, $DO_{x,t}\alpha$, $B_{x,t}p$ and $I_{x,t}p$ are of type *proposition*, where r is of type *untensed proposition*, p is of type *proposition*, t is of type *time*, x is of type *agent*, and α of type *action*. They are read as "r is true at time t," "x did α at time t," "x believes at time t that p," and "x intends at time t that p," respectively.

In the rest of this section we offer a simple account of beliefs and declarative sentences in a default theory. Intentions are added to the account in the next section.

Beliefs are constrained by both axioms and default rules. We assume that an agent's beliefs at any one time follow the standard weak S5 axioms:

The beliefs of one agent at one time are taken to be consistent, distributive over conjunctions, closed under logical consequence and positive introspection. Beliefs need not be true, but should be believed to be.

[Consistency] $\vdash B_{x,t}p \supset \neg B_{x,t}\neg p$

[Closure] $\vdash B_{x,t}p \,\&\, B_{x,t}(p \supset q) \supset B_{x,t}q$

[Positive introspection] $\vdash B_{x,t}p \supset B_{x,t}B_{x,t}p$

[Negative introspection] $\vdash \neg B_{x,t}p \supset B_{x,t}\neg B_{x,t}p$

[Accuracy] $\vdash B_{x,t}(B_{x,t}p \supset p)$

We assume that agents remember their previous beliefs.

[Memory] $\vdash B_{x,t}p \supset B_{x,t+1}B_{x,t}p$

More important is what beliefs x has about p at t+1. Agent x could continue to believe p at t+1 if he believed it at time t. He could also come to believe p if he believes that some other [reliable] agent y believes it. Obviously, these two rules could come into conflict, which could be resolved by giving priority to x's previous belief or to x's belief about y's belief, or by allowing different extensions for different possibilities. For the moment we opt for the former.

[Persistence] $\vdash B_{x,t+1}B_{x,t}p \supset B_{x,t+1}p$

The belief that an action has been performed is acquired from observation of an action. The observability axiom is oversimplified as we take agents to be observed, rather than actions. Observation one agent is assumed to imply observation of all actions performed by him.

[Observability] $\vdash DO_{x,t}\alpha \,\&\, DO_{y,t}Obs(x) \supset B_{y,t+1}DO_{x,t}\alpha$

As rules of inference we assume modus ponens and the usual rule of necessitation:

[Necessitation] If $\vdash p$ then $\vdash B_{x,t}p$

Finally, two default rules are necessary. They both address how new beliefs can be added to old ones. The first allows for transfer of beliefs from one agent to another, provided that the new beliefs are consistent with the old ones.

[Belief transfer rule] $B_{x,t}B_{y,t}p \Rightarrow B_{x,t}p$

The second default associates with the utterance of sentences bearing particular linguistic features an aspect of the speaker's mental state; for declarative sentences, for example, that he believes the propositional content.

[Declarative rule] $DO_{x,t}(p.) \Rightarrow B_{x,t}p$

Because the relation between an utterance and the attitude it expresses is given as a default, and is thus defeasible, there is no need to give anything more complicated. As we shall see, the inference of mutual belief follows from the interaction of these rules.

While the belief transfer rule allows simple (i.e., less deeply nested) belief formulas to be inferred from complex ones, the observation axiom allows complex formulas to be inferred from simple ones. The tension between the application of these two principles is crucial to enabling lies to function much like true assertions, while differing in that liars, for example, are not convinced by their own lies.

One more mechanism is necessary: we need, for example, to allow an agent x, who believes that $DO_{y,t}(p.)$ and that $B_{y,t}p$ is a *default* consequence of $DO_{y,t}(p.)$, to come to believe that $B_{y,t}p$, as long as $B_{y,t}p$ is consistent with x's other beliefs. It is necessary, therefore, that the beliefs of agents be closed under the default rules. One way to view this mechanism is as an analogue for defaults of closure under logical consequence. We need something like the following:

$B_{x,t}p \;\&\; B_{x,t}(p \Rightarrow q) \Rightarrow B_{x,t}q.$

One problem here, of course, is that such an expression is ill-formed, both as a formula in the base language and as a default rule. But, more importantly, it does not capture the fact that it is the *totality* of x's beliefs that must be closed under defaults, not just those that follow from some p. We account for this requirement by adding the following metarule, mirroring the closure of the axiom system under beliefs:

For all agents x and times t, if $p \Rightarrow q$ is a default rule, so is $B_{x,t}p \Rightarrow B_{x,t}q$.

Let W_0 be the set of axioms given above, D_0 the set of default rules.

We can now examine the interaction of axioms and rules in various settings. A bit of notation will be useful. Let (D,W) be a default theory. For any two sets $E, E' \subseteq L$, let

$$D(E,E') = \{\beta \,\|\, \alpha \Rightarrow \beta \in D, \alpha \in E, \neg\beta \neg \in E'\},$$

$$MB_{x,y,t}p = B_{x,t}p \,\&\, B_{y,t}p \,\&\, B_{x,t}B_{y,t}p \,\&\, B_{y,t}B_{x,t}p \,\&\, ...,$$

$$^*B_{x,t}p = MB_{x,x,t}p, \text{ and}$$

$$\phi = DO_{S,0}p. \,\&\, DO_{S,0}Obs(H) \,\&\, DO_{H,0}Obs(S) \,\&\, DO_{S,0}Obs(S) \,\&\, DO_{H,0}Obs(H).$$

It will be convenient at times to treat $MB_{x,y,t}p$, $^*B_{x,t}p$, and ϕ as the set of their conjuncts.

We can now examine our key examples more carefully.

Example 1 (continued). [Sincere assertion] Let S utter a declarative sentence with propositional content p to H at a time 0, where $B_{S,0}p$ and there are no other beliefs about p or believing p. Let it also be the case that S and H are observing each other (and themselves), so that $DO_{u,0}$(Obs v) is true for all substitutions of S and H for u and v. An extension of these assumptions under the the rules D_0 contains the mutual belief by S and H that p, as well as the mutual belief that the observation conditions hold.

Slightly more formally, we can show that the following theory E is an extension of the default theory (ϕ, D_0).

$$E = Th(\phi \cup MB_{S,H,1}(\phi \,\&\, B_{S,0}p \,\&\, p) \cup {}^*B_{S,0}(\phi \,\&\, p) \cup {}^*B_{S,1}(\phi \,\&\, p) \cup {}^*B_{H,1}(\phi \,\&\, p).$$

Before doing so, however, let us examine the rules that contribute to the licensing in E of the first few terms of its most important part, $MB_{S,H,1}p$. Figure 1 illustrates the derivation of the four shortest terms. The lines ending in arrowheads indicate applications of the default rules; those without arrowheads indicate applications of the axioms.

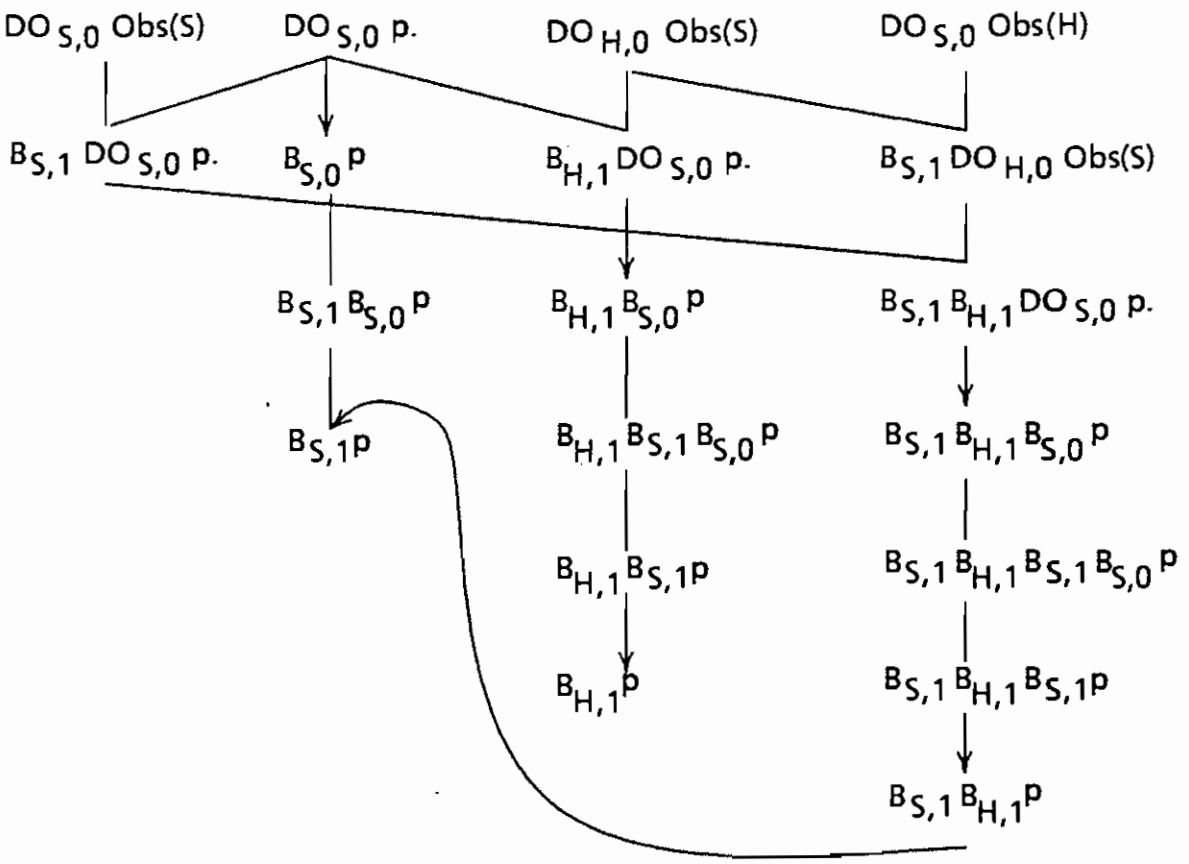

Figure 1. A partial Derivation of the Consequences of a Declarative

To show that E is an extension of (ϕ, D_0), let $E_0 = \phi$ and $E_{i+1} = Th(E_i) \cup D(E_i, E)$. $Th(E_0)$ contains the proposition that everyone comes to believe ϕ and, by repeated application of the observation axiom and closure axioms, that ϕ comes to be mutually believed. The default rule for declaratives lets us add that S initially believes p. Thus

$$E_1 = Th(\phi \cup MB_{S,H,1}\phi \cup {}^*B_{S,1}\phi \cup {}^*B_{H,1}\phi \cup B_{S,0}p).$$

In E_2 we can add that S continues to believe p, by virtue of memory and persistence, that $^*B_{S,0}p$ by repeated application of introspection, and that it becomes mutually believed that S believed p at time 0, by repeated application of the declarative rule.

$$E_2 = Th(E_1 \cup {}^*B_{S,0}p \cup \{B_{S,1}p, B_{S,1}B_{S,0}p\} \cup MB_{S,H,1}B_{S,0}p \cup {}^*B_{S,1}B_{S,0}p \cup {}^*B_{H,1}B_{S,0}p.$$

In E_3, all S's beliefs can be advanced to time 1 by nested applications of memory and persistence.

$$E_3 = Th(E_2) \cup MB_{S,H,1}B_{S,1}B_{S,0}p \cup MB_{S,H,1}B_{S,1}p \cup {}^*B_{S,1}B_{S,1}B_{S,0}p \cup {}^*B_{S,1}B_{S,1}p \cup {}^*B_{H,1}B_{S,1}B_{S,0}p \cup {}^*B_{H,1}B_{S,1}p.$$

Finally, in E_4, we find mutual belief that p, as expected, from repeated application of the transfer rule.

$$E_4 = Th(E_3) \cup MB_{S,H,1}p \cup {}^*B_{S,1}p \cup {}^*B_{H,1}p.$$

E is now $Th(E_4)$, as the defaults can contribute nothing further.

We conjecture that E is the only extension of (ϕ, D_0) and, to some extent, the division of labor between axioms and default rules is designed to ensure this. We return to this question after the second example, in which some of the rules are blocked.

Example 2 (continued). [Lie] This case is much like the previous one, but we now assume that $B_{S,0}\neg p$. The consequences for H should be exactly the same as with a sincere assertion, and so should the consequences for S, except that S continues to believe $\neg p$. The following E is an extension of $(\phi \& B_{S,0}\neg p, D_0)$:

$$E = Th(\phi \cup MB_{S,H,1}(\phi \& B_{S,0}p \& p) - \{B_{S,1}B_{S,0}p, B_{S,1}p\} \cup \{B_{S,1}B_{S,0}\neg p, B_{S,1}\neg p\} \cup {}^*B_{S,1}(\phi \& \neg p) \cup {}^*B_{H,1}(\phi \& p) \cup {}^*B_{S,0}(\phi \& \neg p)).$$

First, in Figure 2, we show informally the derivation of the first few terms of what is mutually believed about p at time 1. The crossed-out terms are those introduced by default rules in the standard case but blocked here.

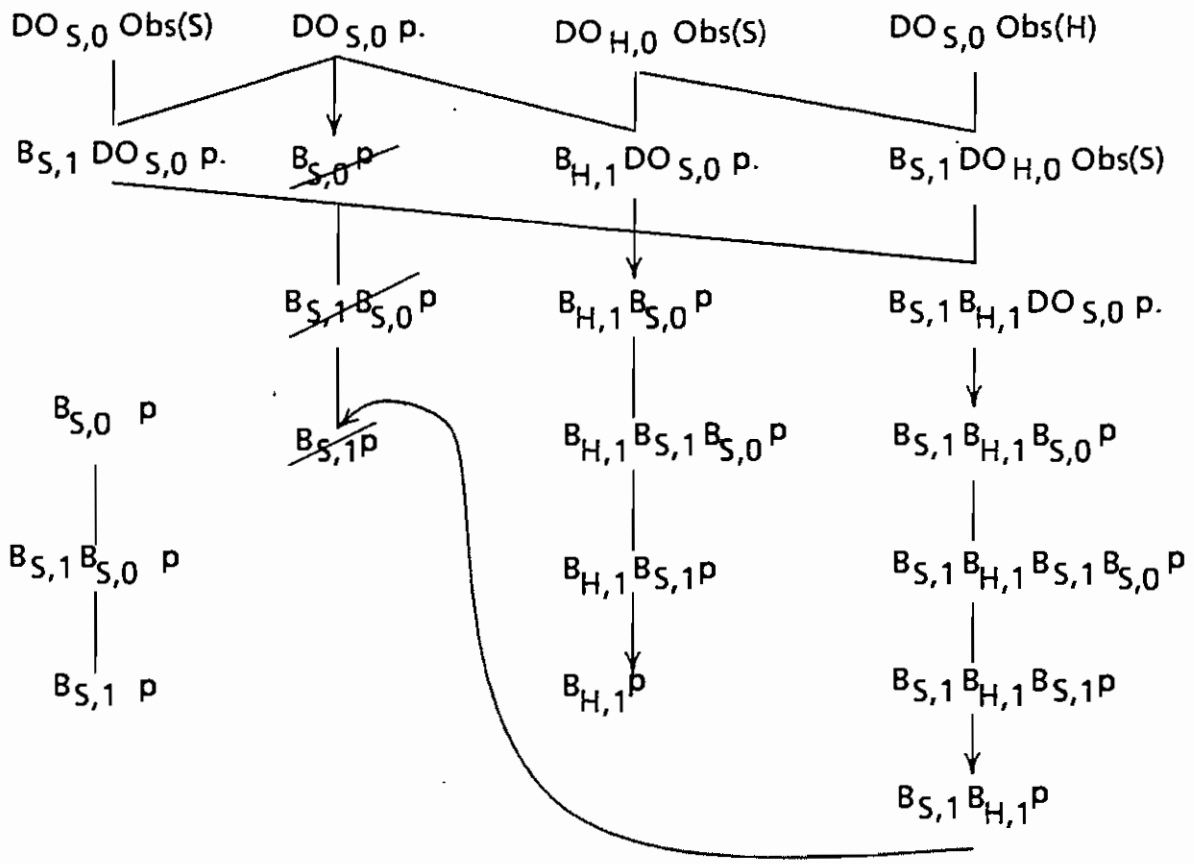

Figure 2. A Partial Derivation of the Consequences of a Lie.

This time we take $E_0 = \phi \ \& \ B_{S,0} \neg p$. E_1 is as in Example 1, except that it does not contain $B_{S,0}p$ because the declarative rule is blocked by $B_{S,0} \neg p$. The last three terms are the result of applying introspection, memory, and persistence to $B_{S,0} \neg p$.

$$E_1 = Th(\phi \cup MB_{S,H,1}\phi \cup {}^*B_{S,1}\phi \cup {}^*B_{H,1}\phi \cup {}^*B_{S,0} \neg p \cup {}^*B_{S,1}B_{S,0} \neg p \cup {}^*B_{S,1} \neg p).$$

In E_2 it becomes mutually believed that S believed p at time 0, by repeated application of the declarative rule, with the exception that S does not change his mind about $\neg p$.

$$E_2 = Th(E_1 \cup MB_{S,H,1}B_{S,0}p - \{B_{S,1}B_{S,0}p\} \cup {}^*B_{H,1}B_{S,0}p).$$

In E_3, memory and persistence apply to all S's beliefs at time 0:

$$E_3 = Th(E_2) \cup MB_{S,H,1}B_{S,1}B_{S,0}p - \{B_{S,1}B_{S,1}B_{S,0}p\} \cup MB_{S,H,1}B_{S,1}p - \{B_{S,1}B_{S,1}p\} \cup {}^*B_{H,1}B_{S,1}p.$$

In E_4, we confirm mutual belief that p at time 1, save that S continues to believe $\neg p$:

$$E_4 = Th(E_3) \cup MB_{S,H,1}p - \{B_{S,1}p\} \cup {}^*B_{H,1}p.$$

No further rules apply and $E = E_4$.

Conjecture: E is the only extension of $(\phi \& B_{S,0} \neg p, D_0)$.

Other defective cases of the use of declaratives can be handled similarly. Initial assumptions are not overridden by beliefs about the content of the declarative, but mutual belief about the existence of the utterance and the observation facts is achieved, along with as much of mutual belief regarding the propositional content as is consistent with the assumed private facts.

One might ask here whether different belief revision strategies can be described in Default Logic. One obvious step to try is to change some of the axioms into default rules. A belief revision strategy in which old beliefs do not always persist can be obtained by replacing the persistence axiom with the persistence rule $B_{x,t+1}B_{x,t}p \Rightarrow B_{x,t+1}p$. Thus, if S declared that p in a state 0, where $B_{S,0}B_{H,0} \neg p$, the axioms would let us infer $B_{S,1}B_{H,1}B_{H,0} \neg p$ (that H continues to believe that he used to believe $\neg p$) and $B_{S,1}B_{H,1}B_{S,1}p$ (that H comes to believe that S believes p). The question is, what, at time 1, S's beliefs about H's beliefs about p ought to be. Having both persistence and belief transfer rules would ensure two extensions of the set of assumptions: one in which $B_{S,1}B_{H,1} \neg p$ (i.e., in which H's beliefs are believed to persist) and an [incompatible] one in which $B_{S,1}B_{H,1}p$ (i.e., in which H is believed to have been convinced by the declaration.) The theory would then give no precedence to either.

One might wish, intuitively, for a single extension in which $B_{S,1}(B_{H,1} \neg p \vee B_{H,1}p)$, i.e. in which S would not be committed to believing one outcome rather than the other. I do not believe this can be done in Default Logic with the current definition of extension.

Adding Intentions

The main characteristic of illocutionary acts is that they are performed successfully only when the hearer recognizes some of the speaker's intentions. In the case of a sincere assertion, the hearer should recognize (i.e., come to believe) that the speaker intended to make the utterance, that he intended that the hearer recognize that he believes the propositional content p, that he intended that the hearer recognize that he intended that the hearer recognize that he believes the propositional content p, and so on. Schiffer [1972] provides a detailed discussion of these conditions. In this

section, we extend our analysis to include the role of the intentions the speaker has about his utterances and their consequences for the mental state of the participants. As with beliefs, we propose an analysis in which the speaker's intentions, including those that have to do with the hearer's recognizing his intentions, are assumed as the result of the application of default rules, and depend both on the form of the actions and on the facts about the joint mental state of the participants. A liar who initially believes $\neg p$ cannot intend to believe p after his declarative. Under our model of belief, if S believes that H believes $\neg p$, then S cannot intend that H will come to believe p, at least not as the result of S's declaring p.

The formal analysis of the full notion of intention and its relation to desires, goals, and beliefs is difficult and still quite controversial (e.g., Bratman [1984], Cohen and Levesque [1987a]). The version given here focuses on the interaction between intentions and beliefs; it is assumed, moreover, that the objects of intentions are [tensed] propositions. Intentions are taken to be consistent – an agent cannot intend that both p and $\neg p$ will hold simultaneously.

[I-consistency] $\vdash I_{x,t}p_{t'} \supset \neg I_{x,t}\neg p_{t'}$

We also assume that intentions are consistent with beliefs, so that believing that p cannot become true is inconsistent with rationally intending that it should become true.

[IB-consistency] $\vdash I_{x,t}p_{t'} \supset \neg B_{x,t}\neg p_{t'}$

We add a new default rule and reformulate another slightly. First we assume that actions are taken to be performed intentionally, as long as this is consistent.

[Intentionality] $DO_{x,t}\alpha \Rightarrow I_{x,t}DO_{x,t}\alpha$

The declarative rule can now be reformulated so that only intentionally performed declaratives are taken to indicate that the speaker believes the propositional content.

[Declarative rule] $I_{x,t}DO_{x,t}(p.) \Rightarrow B_{x,t}p$

These new axioms and rules are relatively uncontroversial. More difficult is the question of the closure of intentions under logical consequence and the default rules. Cohen and Levesque [1986] present two different forward-looking attitudes, GOAL and P-GOAL, and argue that only GOAL is closed under logical consequence. A full integration of their analysis (or of something similar), within the framework presented earlier is beyond the scope of this paper. We assume analogues for intentions of the closure rules we used for beliefs. First, an axiom for logical closure:

[I-closure] $\vdash I_{x,t}p \,\&\, B_{x,t}(p \supset q) \supset I_{x,t}q$

We also assume that intentions are closed under the default rules, implemented as before through a metarule.

> For all agents x and times t, if $B_{x,t}p_{t'} \Rightarrow B_{x,t}q_{t'}$ is a default rule, so is $I_{x,t}p_{t'} \Rightarrow I_{x,t}q_{t'}$.

Let D_1 be the set of default rules obtained by adding the new rules to those given in the previous section then closing the class under the metarule for belief and that for intention.

Although as a general analysis, our view of intention is controversial, we need to deal here only with an agent's intentions relative to the speech act he is in the process of performing and its resulting state. The aspect of intention that is especially crucial to the analysis of speech acts is the commitment to – or the willingness to accept – the consequences of an utterance. What needs to be captured is the fact that, if an agent performs an utterance expressing some aspect of his mental state, say, a belief, and if he believes he is being observed, then a default consequence is that he intends to have it recognized that he holds the belief. This must, of course, be a default consequence, since there may be prior reasons for believing that he, in fact, does not in fact hold the belief.

Example 1 (revisited). Let us consider now the extension E' of the set of assumptions of Example 1 under the revised set of defaults D_1. All the consequences E under the old set of defaults D_0 are in E'. The extension E' now also contains $I_{S,0}q$, $B_{S,1}q$, and $B_{H,1}q$ for every formula q in E'. Thus, the default consequences include the fact that the utterance was performed intentionally ($I_{S,0}DO_{S,0}(p.)$), that it was intended to be recognized as such ($I_{S,0}B_{H,1}I_{S,0}DO_{S,0}(p.)$), and that this be recognized ($I_{S,0}B_{H,1}I_{S,0}B_{H,1}I_{S,0}DO_{S,0}(p.)$), and so on. S also intends that the perlocutionary effects take place (e.g., $I_{S,0}B_{H,1}p$), that this intention be recognized ($I_{S,0}B_{H,1}I_{S,0}B_{H,1}p$), and so on.

Here too, the case of lies is only slightly different. Initially S believes $\neg p$, and thus does not intend that he should not believe $\neg p$, but, as before, S intended to perform the utterance, intended that the hearer H recognize his intention to utter something, intended that H believe p, and that H recognize his intention that H believe p. It must be assumed that it is not the case that S intends that H should believe that S should believe $\neg p$: the contrary leads to exposure of the liar.

What is a Speech Act?

Speech acts can be defined in our framework by stating conditions that must hold in the extension(s) obtained by closing under the default rules D_1 the theory describing the initial mental state of the participants, the fact of the utterance, and the conditions of observation. We attempt here to define the acts of asserting, informing, lying, and convincing.

Searle points out that illocutionary acts are somewhat different when examined from the perspectives of the speaker and hearer, respectively. An illocutionary act has been performed *successfully* if the speaker did it while in a certain mental state, whereas it has been *fully consummated* if the hearer recognized that the speaker performed it successfully. The fully consummated act is what Austin calls the *securing of uptake*. Both perspectives are available in our analysis; examples are given below. Perlocutionary acts are defined in terms of the change in mental state brought about by the utterance.

One of the essential features of a theory of illocutionary acts is its account of the relation between explicit performative and declarative sentences. Some accounts (e.g., Searle [1969]) do not attempt to relate them at all. We believe that the theory should predict that an utterance of "I hereby assert that p" is indeed an assertion if and only if the speaker's mental state when making the utterance satisfies the conditions of the action's being an assertion, by virtue of the utterance of a declarative sentence. Such an account requires a more detailed analysis of the propositional content of the utterance than we can provide (but see Cohen and Levesque [1987b]). Asserting p is doing something that is intended to bring about recognition by the hearer that the speaker believes that p, and doing so sincerely and overtly. S *successfully asserted* p to H in performing an action (utterance) U at time 0 if $B_{S,0}p$ & $I_{S,0}B_{H,1}B_{S,0}p$ & $I_{S,0}B_{H,1}I_{S,0}B_{H,1}B_{S,0}p$ & ... This definition allows S to not intend that H actually come to believe p: the utterance of a declarative sentence in a situation in which it is mutually believed that H believes $\neg p$ is still an assertion according to this definition, and S cannot intend that H believe p in that circumstance. Lies, however, are not assertions, as they are not sincere, nor are any actions by S that are not done overtly. S *fully consummated* an assertion that p if S asserted that p and H recognized that S did so.

Another possible definition of asserting is that S should intend that S and H come to mutually believe that S believes p. This definition would be too weak if we were not already requiring that S be the agent of the action: it would then be satisfied by having a reliable third party declare p to S and H simultaneously. Even if we require S to be the agent, if actions need not be intended, S and H could mutually believe that S believes p if S did something overtly and unintentionally (say, move a block)

and perceived its consequences, say, that the block is at location L. We would not want to say that any assertion had been performed under the circumstances. It should be noted that both defining conditions stated here (and many others) are true in the extension of the default theory given in Example 1, but not true in that of Example 2.

Assuming our definition of assertions, it is quite simple to distinguish them from lies and convincings. Lies are insincere assertions: S *lied about* p to H in performing an action (utterance) U if $B_{S,0}\neg p$ & $I_{S,0}B_{H,1}B_{S,0}p$ & $I_{S,0}B_{H,1}I_{S,0}B_{H,1}B_{S,0}p$ & ... Illocutionary acts, consequently, are those resulting in a state in which the speaker has a certain attitude and overtly intends that it be recognized. In perlocutionary acts, the overtness is unnecessary, as with convincing: S *convinced* H of p in performing an action (utterance) U if S uttered U, intending that H believe p, and H comes to believe p.

Informing is asserting with the intention that the content be believed. S *informed* H that p in performing an action (utterance) U if $B_{S,0}p$ & $I_{S,0}B_{H,1}(p$ & $B_{S,0}p)$ & $I_{S,0}B_{H,1}I_{S,0}B_{H,1}(p$ & $B_{S,0}p)$ & ...

Irony and Indirection

One of the main reasons for basing an analysis of illocutionary acts on changes in mental state is to circumvent the problems created by accounts that postulate a direct connection between utterance features and illocutionary act type – in particular the treatment of nonserious and indirect illocutionary acts. So far, we have attempted to describe the effect of utterances on mental state, on the basis of their features and the conditions of utterance. Let us now examine the nonserious and indirect uses a bit more carefully.

Propositional attitudes play three quite distinct roles in speech act theory; accordingly, these are clearly distinguished from each other in our treatment. First, sentences (utterances, actually) are conventionally related by their features to attitudes that the speaker could, but need not, have towards, among other things, the world and subsequent actions of both speaker and hearer. Declaratives are related to beliefs of the speaker, requests to intentions or desires with regard to actions of the hearer, commissives to commitments of the speaker to future actions, and so on. Let us call these the *literal attitudes* carried by the utterances. In our framework, the relation between the features of an utterance and the literal attitude it expresses is captured by the declarative default rule (and its analogues for other clusters of utterance features and attitudes.)

Second, a speaker makes use of utterances to express attitudes she may or may not have towards the world and subsequent actions of speaker and hearer. Although

these attitudes may correspond to the literal attitudes carried by the utterances the speaker makes to express them, they need not, as in indirect and non-serious ues, not need they be held by the speaker, as in insincere uses. Let us call these attitudes of the speaker the *core attitudes*.

Finally, besides core attitudes, speakers have attitudes towards the utterances themselves and towards the change in the mental states of all participants that is effected by the recognition of the speaker's intentions. Let us call these *Gricean attitudes*. In our framework, expressions describing the Gricean attitudes are of the form $I_{S,0}B_{H,1}q$, $I_{S,0}B_{H,1}I_{S,0}B_{H,1}q$, etc., where q is a core attitude. They (or rather their descriptions) are generated by application of the various axioms and default rules to the assumptions regarding the initial mental state of the participants and the action performed.

Gricean attitudes characterize the class of illocutionary acts. A speaker performs a communicative act successfully if he has the right Gricean attitudes towards the core attitude he is trying to convey, as discussed in the definitions of asserting and informing presented in the preceding section. The Gricean attitudes are parameterized, if you will, by core attitudes.

As Searle [1975] suggests, communicative conventions can be systematically violated to good effect. The violations take several forms. The most obvious kind occurs when it is incompatible with the initial mental state of the participants that the speaker should have the Gricean attitudes expressed by the given utterance. This is the case with ironic usage. After a terrible meal shared with the hearer, the speaker, in saying "This is the best meal I ever had," does not really believe that the meal was wonderful, nor does he intend that it should be recognized that he does; the initial joint mental state is such that all this is mutually believed. Thus, no illocutionary act with core attitude $B_{S,0}$(this is the best meal) can be performed by uttering this sentence. However, this need not mean that *no* useful illocutionary act was performed.

What the speaker of an ironic utterance *did* mean can be determined by finding a related core attitude that the speaker could have Gricean intentions about. In ironic uses, the propositional attitude in the core remains the same, but the propositional content changes – in this case, to something like "this is the worst meal I ever had." The speaker's belief in that proposition can now be the core attitude for a successful illocutionary act, as it is possible now for the speaker to believe that this was the worst meal, intend that that intention be recognized, and so on. The "derived" content is new to the hearer and it is consistent with what he knows of the speaker's previous mental state that the speaker should hold it. While it still remains, of course, to establish the details of the process for determining the conveyed core attitude from the literal one, we have nothing new to say about that process here. Note that ironic uses can also be insincere: the speaker may in fact believe privately

that he has just had a pretty good meal yet convey ironically that it was terrible. He can also be recognized as lying, etc.

Indirect uses fall in two categories. The first, exemplified by "It's cold here," said by one person to another upon entering a cold room. has the speaker expressing a true attitude, but his expressing it cannot lead to a change in the beliefs of the hearer, as the fact being expressed is already mutually believed (or can be presumed to be mutually believed by ideal interlocutors.) The second kind of indirect use is exemplified by a different use of the same sentence, uttered, say, by a tenant to his landlord over the telephone. Here again the literal core attitude is sincerely held, but this time the literal assertion is a useful illocutionary act: the speaker intends that the hearer recognize his belief, recognize also that he intends that that intention be recognized, and so on. However, the literal core attitude does not exhaust the message the speaker intends to convey.

In the indirect cases, the derived content can be obtained from inferences regarding the speaker's plans. Heuristic methods for plan recognition have been presented by Schmidt et al. [1979] and Allen and Perrault [1980]. A formal definition of the problem based on an extension of circumscription is outlined by Kautz and Allen [1986]. Here too, it must be noted that the result of the plan inference – e.g., that the speaker wants the hearer to turn the heat up – must be appropriate as the core attitude of a successful illocutionary act. In other words, it must be possible for the speaker to have the appropriate Gricean intentions towards the derived attitude. If, for example, it was mutually believed that there was no heat to be turned up in the building, the speaker could not be trying to convey, even indirectly, his desire to have it be turned up. She might, however, be trying to imply that a furnace should be installed. Also, as was the case with irony, the speaker can be insincere about the indirectly conveyed attitudes: he may not really want the heat turned on, but merely wish to distract the landlord from watching the Super Bowl. In fact, indirection and irony can be used simultaneously. In saying "It's cold here" with an ironic intonation or in a very warm room, the speaker can be indirectly conveying that he wants the room to be made colder. And, of course, he may be lying about that too.

Finally, a comment about force indicators other than sentence type. Some violations of conventional uses of sentence type can be marked, in particular by gestures and intonation. Nothing in our treatment precludes, at least in principle, the possibility that the rules relating utterance features to literal attitudes could be sensitive to information conveyed by other markers. Oversimplifying, if the derived core attitude in an ironic declarative with propositional content p were $B_{S,t} \neg p$, then a feature rule could be added expressing directly that, if S utters a declarative with propositional content p under ironic intonation, it should be assumed that $B_{S,t} \neg p$. One would probably want this rule to have precedence over the declarative rule, in the sense that, if it could apply, the other should not. Some problems arise in default logic if

rules with distinct but not mutually exclusive defaults are allowed. These are discussed by Reiter and Criscuolo [1983] and by Etherington [1986].

Comparison with Cohen and Levesque [this volume]

Cohen and Levesque's paper in this volume (hereafter C&LTV) is a major revision and elaboration of C&L, developed largely in parallel with this one. It is a more ambitious enterprise than this one, containing a comprehensive theory of intention and derived explanation of directives, all developed in a monotonic famework. My more modest aims are to explore the relation between action, observation, and belief revision. Some differences between the two acounts are worth noting.

C&LTV's account allows participants to change their mind about asserted propositions (i.e. go from believing $\neg p$ to believing p), whereas mine does not. This is obviously desirable in some circumstances, but the question is what price must be paid to allow it. Their version of our declarative default is an axiom postulating that the hearer, observing a declarative, comes to believe the core attitude, namely that the speaker believes p, as long as the hearer does not believe that the speaker is insincere about believing p. Simplifying their formulation slightly, and recasting it in the notation of this paper, they assume

$$B_{H,0}(DO_{H,0}(\text{Obs S})) \& B_{H,0}(DO_{S,0}(p.)) \& \neg B_{H,0} \neg \text{Sincere}_{S,H,0} B_{S,0} p \supset B_{H,1} B_{S,1} p$$

More deeply nested forms are also postulated but are not immediately relevant here. This allows an agent to go from believing $\neg p$ to believing p, and to go from not believing p to believing it, neither of which is allowed by the present account.

The price they pay for allowing true changes of belief is twofold. First, they need to postulate an extra concept, sincerity, to gate between H's beliefs before and after an utterance. Second, they make no claims about what happens when the gating condition on sincerity is not satisfied. If they did, say by claiming that H continues to believe whatever he believed before concerning S's beliefs about p, they would make it clear that their solution, like all monotonic accounts of theories of actions, would encounter what McCarthy called the qualification problem: that the only way of being able to prove anything about the state resulting from an action is to be able to prove the truth of all the qualifying preconditions, and that these can be arbitrarily complex.

The notion of sincerity itself is quite tricky, and I am still not convinced that it is necessary. The difficulty is to make it non-trivial, so that using the declarative axiom is not simply a matter of postulating, arbitrarily, that insincerity fails. It should be possible, for example, to prove, in some circumstances, that the failure of

sincerity follows from interesting features of the utterance context, or that it is possible to infer the speaker's insincerity from what he says. C&LTV define sincerity as follows: An agent x is *sincere* with respect to a proposition p towards another agent y if whenever x has the goal to do something that would bring about that y should come to believe p, then x has the goal that y should come to know p, i.e., x has the goal that p should be true after his action and that y should believe it. One difficulty with this notion is that it does not connect in the right way with beliefs at the time of utterance. In Example 3 above, the case of the unsuccessful lie, we argued that it was rational for H to believe that S believes $\neg$p, both before and after S's declarative utterance of p. My understanding is that, in their semantics, it is consistent for H to hold these beliefs *and* to believe that S is not sincere about believing p. Their theory would then predict that, if H initially believes S to not be sincere, H comes to believe that S believes p, contrary to our intuitions that H should come to believe that S believes $\neg$p. The initial condition is not sufficient to block the application of Cohen and Levesque's version of the declarative axiom.

Conclusion

We have argued that a theory of speech acts intended to consider various nonstandard uses of utterances must be based not only on static accounts of the attitudes, but on attitude revision as well. We used Default Logic for a restricted analysis of belief revision, showing how it could be used to account for various insincere, non-serious, and indirect utterances of declarative sentences.

Although speech act theory is typically considered a part of natural language pragmatics, we prefer to think of it as a part of semantics, albeit not a truth-functional one. The restriction of the realm of semantics to truth-conditional aspects of meaning precludes a uniform semantic treatment of declarative, imperative and interrogative sentences, and even leaves out [declarative] explicit performative sentences. Although very useful work has been done on the satisfaction conditions of interrogatives and imperatives, these accounts are unsatisfactory in that they postulate one kind of object as the interpretation of declarative sentences (truth values or intensions or relations between a discourse situation and a described situation), other kinds for questions and imperatives. Making the object of the interpretation a bit more complex allows a uniform treatment. There are two possibilities. The first is what we called the core attitude, while the second is the function from mental state to mental state resulting from the performance of the utterance. If the main argument of this paper is correct, systematic handling of explicit performatives, irony and indirection require the second, more complex, route. I also believe it to be the only way leading to a systematic semantics of extended discourses, but that question is best left for another time.

Acknowledgments

The research reported here was made possible in part by a gift from the Systems Development Foundation. Alex Borgida, Phil Cohen, Robin Cohen, David Etherington, Hector Levesque, Jerry Morgan, Martha Pollack, and Ray Reiter kindly read and commented on earlier drafts, but any remaining errors are mine alone.

References

Allen, J.F., and Perrault, C.R., Analyzing intention in utterances. *Artificial Intelligence*, 15, 143-178, 1980. Reprinted in Grosz et al. (eds.), 1986.

Austin, J.A., *How to Do Things with Words*. New York: Oxford University Press, 1962.

Bratman, M., Two faces of intention. *Phil. Review*. 93, 3, 375-405. 1984.

Bruce, B.C., Generation as social action. In *Theoretical Issues in Natural Language Processing*, Urbana-Champaign: Assoc. for Computational Linguistics, 64-67, 1975. Reprinted in Grosz et al. (eds.), 1986.

Cohen, P.R. and Levesque, H., Speech Acts and Rationality. *Proc. of 23rd Annual Meeting of Assoc. for Computational Linguistics*, 49-59, 1985.

Cohen, P.R. and Levesque, H., Persistence, intention and commitment, to appear in P. Cohen, J. Morgan, and M. Pollack (eds.), *SDF Benchmark Series: Plans and Intentions in Communication and Discourse*, Cambridge, MA: MIT Press, 1987a.

Cohen, P.R. and Levesque, H., Rational interaction as the basis for communication, to appear in P. Cohen, J. Morgan, and M. Pollack (eds.), *SDF Benchmark Series: Plans and Intentions in Communication and Discourse*, Cambridge, MA: MIT Press, 1987b.

Cohen, P.R. & Perrault, C.R., Elements of a plan-based theory of speech acts. *Cognitive Science* 3, 177-212, 1979. Reprinted in Webber and Nilsson, 1981, and in Grosz et al., 1986.

Etherington, D.W., *Reasoning with Incomplete Information: Investigations of Non-Monotonic Reasoning*, Ph.D. Thesis, Dept. of Computer Science, University of British Columbia, 1986.

Grice, H.P., Meaning. *Phil. Review*. 1957.

Grosz, B.G., Sparck Jones, K. and Webber, B.L., *Readings in Natural Language Processing*. Los Altos: Morgan Kaufmann, 1986.

McCarthy, J., Circumscription – a form of non-monotonic reasoning. *Artificial Intelligence.*, 13, 27-39, 1980. Reprinted in Webber and Nilsson (eds.), 1981.

Perrault, C.R. and Allen, J.F., A plan-based account of indirect speech acts, *Am. J. of Computational Linguistics*, 1980.

Reiter, R., A logic for default reasoning. *Artificial Intelligence*, 13, 81:132 1980.

Reiter, R. and Criscuolo, G., Some representational issues in default reasoning, *Int. J. of Computers and Mathematics*, 9, 1, pp 1-13, 1983.

Schiffer, S.R., *Meaning*. London: Oxford University Press, 1972.

Schmidt, C.F., Sridharan, N.. and Goodson, J.L., The plan recognition problem: an intersection of artificial intelligence and psychology, *Artificial Intelligence*, 10, 1, 1979.

Searle, J.R., *Speech Acts*. New York: Cambridge University Press, 1969.

Searle, J.R., Indirect speech acts, in Cole and Morgan (Eds.), *Syntax and Semantics, vol 3: Speech Acts*, New York: Academic Press, 1975.

Smith, N.V., *Mutual Knowledge*. New York: Academic Press, 1982.

Strawson, P.F., Intention and convention in speech acts. *Phil. Review*, 73, 4, 1964. Reprinted in Searle, J.R. (ed.), *The Philosophy of Language*, London: Oxford University Press, 1971.

Webber, B.L. and Nilsson, N.J. (eds.), *Readings in Artificial Intelligence*. Palo Alto: Tioga Press, 1981.