Chapter 10
Society

The phrase aroused my interest because of its enigmatic quality: "He brought in a second actor." I stopped; I found that the subject of that mysterious action was Aeschylus and that, as we read in the fourth chapter of Aristotle's Poetics, *he "raised the number of actors from one to two" ... With the second actor came the dialogue and the indefinite possibilities of the reaction of some characters on others. A prophetic spectator would have seen that multitudes of future appearances accompanied him: Hamlet and Faust and Segismundo and Macbeth and Peer Gynt and others our eyes cannot yet discern.*

Jorge Luis Borges, "The Modesty of History" in *Other Inquisitions*

This final chapter deals with commonsense knowledge about the interrelations and interactions among agents. In this domain more than any other, commonsense understanding far outstrips any existing formal model in richness and sophistication. Such an understanding is vital for story understanding, in automated teaching and in intelligent interactive systems. In this chapter, we discuss some issues in human and social interactions that have been studied in the AI literature: common knowledge, plan interaction, communication, ethical values, and possession.

The last three of these issues — communication, ethics, and possession — share a number of characteristics. All three issues arise continually in life in a complex society. All are virtually universal among human societies[1] and, apparently, appear in rudimentary form among some nonhuman species. All three are determined, at least to

[1] There are some nomadic hunting societies that carry virtually nothing with them on their travels. Property is an irrelevant concept to these. As far as I know, all societies have some language and some concept of the permitted and the prohibited.

some degree, by social convention,[2] but within the society they can often be extremely rigid and inescapable. It is easy to develop a representation for any of these issues that will cover the simplest cases; however, giving a deep theory of any raises profound and unresolved philosophical controversies.

10.1 Common Knowledge

Suppose that Tom and Huck are hiding together, and they witness Joe murder Dr. Robinson. Then Tom, Huck, and Joe each know the proposition "Joe murdered the doctor." Moreover, Tom and Huck each know the proposition "Tom, Huck, and Joe each know that Joe murdered the doctor"; Joe, however, does not know this. Further, Tom and Huck each know the proposition "Tom and Huck each know that Tom and Huck each know that Joe murdered the doctor," and so on *ad infinitum.* In short, Tom and Huck have a complete understanding between them about the fact.

The technical term for this is *common knowledge:* Tom and Huck have common knowledge that Joe murdered Dr. Robinson. To represent this, we introduce the syntactic operator "common_know(SAA, P)," meaning that the set of agents denoted by string SAA have common knowledge of sentence P. (A third temporal argument can be added when necessary.) Thus, "Tom and Huck have common knowledge that Joe murdered Dr. Robinson" is represented

common_know(≺
{Tom, Huck}≻, ≺$\exists_I$ occur(I,murder(joe, robinson))≻)

It is possible for every pair of agents in a set to have common knowledge of a fact without there being common knowledge of the fact over the set as a whole. For example, if A and B share a secret ϕ, and each independently tells it to C, then each of the sets $\{ A, B \}$, $\{ B, C \}$, and $\{ A, C \}$ has common knowledge of ϕ, but the set $\{ A, B, C \}$ does not. For example, A does not know that B knows that C knows ϕ. (See Exercise 1 for an example where this distinction makes a big difference.)

Unlike simple knowledge, common knowledge is opaque in the agent argument[3]; the members of a set may share knowledge under one description of the set but not another. For example, the secret hand-

[2] In defense of the statement that ethical values or property are to some degree a matter of social convention, let me point out that I am including in these categories such rules as "Don't drive through red lights" or "A person is considered to have paid a bill as of the moment the biller receives his check, not when the check clears."

[3] Lesperance [1989] argues that the same should be true of single agents.

shake of the Masons is (I presume) common knowledge among the Masons: All Masons know it, know that all other Masons know it, know that other Masons know that all Masons know it, etc. Nonetheless, it does not follow that if Joseph and Dominic are both Masons, then Joseph knows that Dominic knows the secret handshake; Joseph may not know that Dominic is a Mason. Another example: In Renaissance Venice, the Council of Three were secretly chosen; they met masked and did not know each other's identity (according to Mark Twain). They presumably shared a large body of common knowledge. The same three individuals (being prominent Venetians) undoubtedly knew each other, and may, indeed, have shared common knowledge in mufti. The two bodies of common knowledge, however, were quite separate. In the first case, each of them knew that all the members of the Council of Three knew the facts; in the other case, each of them knew that each member of the set { Leonardo, Giuseppe, Pietro } knew the facts. Thus, in a modal theory, the agent argument is an opaque argument; in a syntactic theory, it is a string that denotes a set of agents; in a possible-worlds theory, it is a fluent ranging over sets of agents, whose value may vary from one possible world to another.

Table 10.1 shows some basic axioms of common knowledge in a syntactic language.

Axioms CK.1 and CK.2 establish the basic properties of common knowledge: If agents AA have common knowledge of ϕ, then they all know ϕ, and they all know that they all know ϕ, and so on. Axiom CK.1a, that common knowledge is veridical, is a direct consequence of CK.1 and KNOW.3, that knowledge is veridical. Axioms CK.3 and CK.4 are the usual consequential-closure principle, analogous to axioms KNOW.1 and KNOW.2 of knowledge. Axiom CK.5 states that, if agents AA have common knowledge of ϕ and agents BB have common knowledge that BB is a subset of AA, then agents BB have common knowledge of ϕ. We can use this rule to justify inferences like the following: If Masons have common knowledge of the Masonic handshake, and Joseph and Dominic have common knowledge that they are Masons, then Joseph and Dominic have common knowledge of the Masonic handshake.

Finding an appropriate set description for a set of agents with common knowledge can raise substantial difficulties. For example, consider three strangers who are standing together and together witness a rabbit eating a carrot. Then it is clear that the three people have common knowledge that the rabbit ate the carrot, but under what description of one another? One solution would be to say that the common knowledge is shared among "the people here now," but that involves positing a common language of indexicals among agents. It

Table 10.1 Axioms of Common Knowledge

CK.1. If P is common knowledge among SAA, then every agent in SAA knows P.

[common_know$(SAA, P) \wedge A \in$ denotation$(SAA)] \Rightarrow$ know(A, P).

CK.1a (Veridicality) If P is common knowledge among SAA, then P is true.

common_know$(SAA, P) \Rightarrow$ true(P).

CK.2. (Positive introspection) If P is common knowledge among SAA, then it is common knowledge among SAA that P is common knowledge among SAA.

common_know$(SAA, P) \Rightarrow$
common_know$(SAA, \prec$common_know$(!SAA!, !P!)\succ)$.

CK.3. (Consequential closure) Common knowledge is closed under implication.

[common_know$(AA, P) \wedge$
common_know$(AA, \prec\downarrow P\downarrow \Rightarrow \downarrow Q\downarrow\succ)$] $\Rightarrow$
common_know(AA, Q)

CK.4. (Common knowledge of the axioms) Basic axioms are common knowledge.

If P spells out an axiom (of logic, knowledge, time, or whatever), then

$\forall_{SAA}$ common_know(SAA, P)

is an axiom.

CK.5. (Common knowledge among a subset)

[common_know$(SAA, P) \wedge$
common_know$(SBB, \prec\downarrow SBB\downarrow \subseteq \downarrow SAA\downarrow\succ)$] $\Rightarrow$
common_know(SBB, P).

is worthwhile distinguishing situations where this question does not arise, because all the agents involved know each other. We will represent the state in which all the agents in set AA know one another by common names and have common knowledge of this fact using the state type "acquainted(AA)."

To represent common knowledge in a possible-worlds semantics, we use an accessibility relation "ck_acc$(FAA, W1, W2)$," where $W1$ and

$W2$ are possible worlds and FAA is a fluent that ranges over sets of agents in different worlds. FAA thus corresponds to a particular description of some set of agents. The relation ck_acc($FAA, W1, W2$) holds if $W2$ is consistent with everything that is common knowledge in $W1$ to the set of agents described by FAA. The statement that agents AA have common knowledge of ϕ in $W1$ is expressed by asserting that ϕ is true in every accessible world.

10.2 Multiagent Plans

An agent in a society of many agents must consider his neighbors' actions and reactions in forming and evaluating his plans. Interactions among plans can be divided into a number of basic categories:

- *Cooperation*: A set of agents with a common goal cooperate on a common plan. The structure and analysis of such a plan may be considerably more complicated than the analysis of a single-agent plan. In many cases, there will not be any single agent that knows all steps of the plan. Step-by-step validation of the plan, such as we have seen in single-agent plans, therefore becomes largely irrelevant; there is no agent who knows enough to carry out such a validation. Instead, the analysis centers around the ways in which subtasks are divided among the various agents. Another new issue that becomes important in cooperative plans is the communication of relevant information between agents. This type of planning has much in common with the programming of distributed computer systems [Smith and Davis 1981].

- *Influence*: In a plan of influence, the planner tries to get others to perform actions that further his goals. Here, the other agents are analogous to external physical events that the agent is trying to control. The difference is that the agents are governed by psychological and social rules, rather than physical rules. A major aspect of this planning is predicting how another agent will respond to a request, discussed in Section 10.3.2. Section 10.3.3 shows how a simple plan of influence, involving making a request, can be validated.

- *Contingent interaction*: Other agents intend to perform actions, purely for their own purposes, that may aid or hinder the plan of the planner. Again, as in plans of influence, the other agents may be considered as external events that can be affected in a number of ways.

- *Direct conflict*: Another agent specifically wishes to prevent the planner from accomplishing his goal (or vice versa). Here, the planner must assume that the opponent will always perform the action that he (the opponent) considers most obstructive. Such planning often involves an exchange of hostile, destructive, or deceptive actions.

A single plan may involve different kinds of interactions. For example, a group of agents may cooperate on a plan that involves influencing some other agents and conflict with others. There are also many cases that fall on the borderline. For example, the workings of a company are, in some respects, a cooperative activity of the employees; in some respects, an influencing of the workers by the boss; and in some respects, a conflict between the workers and the boss.

Research in representing and reasoning about such plans is currently in a primitive stage, though some intriguing work has been done. Citations are given in the chapter reference list.

10.3 Communication

The most central type of interactions among intelligent agents are communications, also known as *speech acts*. (We use the terms associated with speech — "speech," "speaker," "hearer," and so on — to apply to any mode of communication: speech, writing, telegraphy, etc.) Speech acts may be divided into five categories: *declarative* acts, which convey information, such as stating "Paris is the capital of France"; *interrogative* acts, which ask for information, such as asking "What is the capital of France?"; *imperative* acts, which make a request or issue a command, such as asking "Please go away"; *exclamatory* acts, which express an emotion, such as crying "Alas!"; and *performative* acts, which, by social convention, bring about a condition, such as "I now proclaim you man and wife" or "I hereby cede my rights to Sherwood Forest." These categories are (nearly) mutually exclusive and exhaustive; almost every speech act belongs to exactly one of the categories.

Another well-known categorization associated with speech acts is the division into *locutionary, illocutionary* and *perlocutionary* acts [Austin 1961]. Unlike the previous categorization, these are categories of speech-act *types*, rather than of *tokens*; any given speech-act token may be described in any of these ways. A locutionary speech act describes the act in terms of the physical characteristics of the acts and the symbols used. An illocutionary speech act describes the act in

Table 10.2 Speech Act Categories

Locutionary:	Shouting "I'm mad as hell!" Whispering "Te amo." Writing "John Hancock."
Illocutionary:	Declaring allegiance to the king. Asking for a raise. Proposing a merger between General Soap and Urumchi University Press.
Perlocutionary:	Boring the audience. Converting a student to Marxism. Charming a guest.

terms of its content. A perlocutionary speech act describes the act in terms of the effect on the hearer. Table 10.2 gives some examples of each:

In this section, we will study locutionary and illocutionary descriptions of declarative, interrogative, and imperative speech acts.

10.3.1 Locutionary Descriptions

Locutionary acts are physical actions that produce sequences of signs in speech, writing, sign language, or other physical medium. We introduce the function "speak(P)," mapping a string of phonemes P to the action of speaking P, and the predicate "pronunciation(S, P, L)," meaning that phonemes P are an acceptable pronunciation of string S in language L. Thus, we can express the statements "Humphrey Bogart said 'Here's looking at you'" and "Lincoln spoke the Gettysburg Address" in the forms

$\exists_P$ pronunciation(P, $\prec$Here's_looking_at_you$\succ$, english) $\wedge$
occur(i202, do(bogart, speak(P))).
$\exists_P$ pronunciation(P, gettysburg_address, english) $\wedge$
occur(i624, do(lincoln, speak(P))).

where "gettysburg_address" is a constant denoting the string $\prec$Fourscore_and_seven_years_ago ...$\succ$.

Speech can accomplish communication only to someone who is present to hear it. Since hearing is a perception, we may apply the theory of perception developed in Section 8.7. For example, we can express a rule such as "If hearer AH is within distance d_hear of speaker AS, then he can hear whether AS is speaking and what he is saying" in the following form:

$$[\,\forall_{S\in I}\ \text{distance}(\text{value\_in}(S,\text{place}(AH)),\ \text{value\_in}(S,\text{place}(AS))) \leq \text{d\_hear} \wedge \text{bpc}(AH, B, \text{behavior}(I))\,] \Rightarrow [\,\text{occurs}(I,\text{do}(AS,\text{speak}(P))) \Leftrightarrow \exists_{AS1}\ \text{occurs}(B,\text{do}(AS1,\text{speak}(P)))\,].$$

10.3.2 Illocutionary Speech Acts

An illocutionary description of a speech act characterizes its content. We introduce the event type, "illoc(AS, AH, M, P)", the event of AS communicating P in mode M to AH. (For simplicity, we assume a single speaker and hearer.) In a declarative act, P will be a sentence in a formal language, and M will be the constant "declarative." (Of course, declarative speech acts may be carried out without speaking an entire natural-language sentence. For instance; a question can often be answered in a single word. We assume, however, that the content of any declarative speech act can be expressed as a sentence in a formal language.) In an imperative act, P will be a term in a formal language denoting the action type that the speaker wants the hearer to perform, and M will be "imperative." We will treat interrogatives as special types of imperatives in which the speaker is requesting the hearer to communicate an answer to his question.

For example, the statement "Becky told Amelia that all crows are black" may be represented

$$\text{occur}(\text{i202}, \text{illoc}(\text{becky}, \text{amelia}, \text{declarative}, \prec\!\succ\forall_X\ \text{crow}(X) \Rightarrow \text{black}(X)\prec\!\succ)).$$

The statement "Becky told Amelia where Joseph was" is interpreted as "Becky told Amelia a true statement of the form 'Joseph is at X' for some constant X," and is thus represented

$$\exists_{X,P}\ P = \prec\!\succ\text{value\_in}(\text{start}(\text{i202}),\text{place}(\text{joseph})) = \downarrow X \downarrow\succ \wedge \text{constant}(X) \wedge \text{true}(P) \wedge \text{occur}(\text{i202}, \text{illoc}(\text{becky}, \text{amelia}, \text{declarative}, P)).$$

The statement "Becky asked Amelia to pass the salt" is represented,

occur(i202, illoc(becky, amelia, imperative,
≺do(amelia,pass(salt42))≻))

The statement, "Amelia asked Becky where Joseph was" is interpreted as "Amelia requested that Becky communicate a true sentence of the form 'Joseph is at X,' where X is a constant," which is represented

occur(i202, illoc(amelia, becky, imperative,
≺ { I | $\exists_{X,P}$ occur(I, illoc(becky, amelia, declarative, P)) ∧
P=≺value_in(start(i202), place(joseph))=↓X↓≻ ∧
true(P) ∧ constant(X) }≻)).

(Note: the down arrows surrounding X in the formula above are scoped to the inner string delimiters.) Table 10.3 shows some basic properties of illocutionary acts.

Axioms IL.1 and IL.3 characterize the relation between locutionary and illocutionary acts. They can serve as the basis for executing an illocutionary act as a locutionary act in a task-reduction planner. The functions "meaning_of(S, K)" and "mode_of(S, K)" in these axioms specify the relation between the string spoken and the content communicated; the content of these functions would be specified by a theory of natural-language semantics. Axioms IL.6 and IL.8 state some necessary conditions that declarative and imperative speech acts be sincere. Additional conditions could be added. For example, we could define a declaration of P to be sincere if the speaker both believes P and believes that the hearer will believe that the speaker is sincere. We could define a request of P to be sincere if the speaker desires P and believes that his requesting P will make the hearer more likely to perform P [Searle 1969].

By combining these axioms with the axioms of belief in Chapter 8, we can justify[4] the plausible inference that if A declares ϕ to B, then B will believe ϕ. Assume that A declares ϕ to B. By IL.2, A and B have common knowledge that A has declared ϕ to B. By axiom CK.1, B knows that A has declared ϕ to B. Using rule BEL.14, we can presume that B will apply IL.4 and infer that A believes ϕ, and that he will further apply rule BEL.13, and infer that ϕ is true.

Stronger initial information will support stronger conclusions. For example, assume (i) that A declares ϕ to B; (ii) that A and B have common knowledge that A is speaking sincerely; and (iii) that A and B have common knowledge that A knows whether ϕ is true. Then it

[4]Modulo the limitation that our notation "plausible" doesn't actually refer to any particular theory of plausible reasoning.

Table 10.3 Axioms for Illocutionary Acts

IL.1. token_of(K, illoc(AS, AH, M, P)) $\Rightarrow$
$\exists_S$ token_of(K,do(AS,speak(S))) $\wedge$
P=meaning_of(S, K) $\wedge$ M=mode_of(S, K).
An illocutionary act communicating P in mode M involves speaking a string whose meaning is P and mode is M.

IL.2. [occur(I,illoc(AS, AH, M, P)) $\wedge$
true_in(start(I),acquainted($\{AS, AH\}$))] $\Rightarrow$
common_know($\prec\{@AS@, @AH@\}\succ$,
$\prec$occur(@I@,illoc(@AS@,@AH@,$\downarrow M\downarrow$, !P!))$\succ$,
end(I)).
The occurrence of an illocutionary act is a matter of common knowledge to speaker and hearer.

IL.3. true_in(start(I), acquainted($\{AS, AH\}$)) $\Rightarrow$
[occur(I,illoc(AS, AH, M, P)) $\Leftrightarrow$
common_know($\prec\{@AS@, @AH@\}\succ$,
$\prec\exists_{S,K}$ token_of(@K@,do(@AS@,speak(S))) $\wedge$
@I@=time_of(K) $\wedge$!P!=meaning_of(S, K) $\wedge$
$\downarrow M\downarrow$=mode_of(S, K)$\succ$,
end(I))].
A communication of P from AS to AH occurs just if they have common knowledge that AS spoke some string S that means P.

IL.4. plausible(token_of(K,illoc(AS, AH, M, P)), sincere(K)).
Illocutionary acts are typically sincere.

IL.5. token_of(K,illoc(AS, AH,declarative,P)) $\Rightarrow$ sentence(P).
The content of a declarative act is a sentence.

IL.6. [token_of(K,illoc(AS, AH,declarative,P)) $\wedge$ sincere(K)] $\Rightarrow$
believe(AS, P,start(time_of(K))).
The speaker believes a sincere declarative speech act.

IL.7. token_of(K,illoc(AS, AH,imperative,P)) $\Rightarrow$
sort_of(denotation(P))=event $\wedge$
actor_of(denotation(P))= AH.
The content of an imperative act is a description of an action by the hearer.

IL.8. [token_of(K,illoc(AS, AH,imperative,P)) $\wedge$ sincere(K)] $\Rightarrow$
goal(AS, P,start(time_of(K)))].
If an imperative act is sincere, then the speaker wants the specified action to be carried out.

is possible to infer that A and B have common knowledge of ϕ when the communication is complete (Exercise 3).

In an analogous way, one would like to support the inference that if A requests something from B then, under suitable conditions, B will perform it. Such an inference is necessary to construct plans that involve the cooperation of agents who do not necessarily share the planner's goals. The problem is to spell out the requisite conditions. So far, this has only been done for very idealized microworlds. Section 10.3.3 shows the validation of a simple plan of influence involving a request, using an *ad hoc* assumption that this particular hearer always carry out the requests of the speaker immediately, if possible. (Schank and Abelson [1977] discuss a sequence of request forms, in increasing order of power and decreasing order of ease: asking; invoking a theme relating the two actors (e.g., "We've always been friends"); informing the hearer of the reason for the request; bargaining with the hearer; threatening the hearer; and overpowering the hearer.)

10.3.3 Sample Verification of a Plan of Influence

In this section, we illustrate how a simple plan of influence in a toy world can be verified. This verification combines primitives and rules from Chapters 5, 8, 9, and 10.

Assume that, in the blocks world of Section 5.3, block A is clear and on top of block B and the hand is above them. Fred manipulates the hand. Jack wishes block B to be clear. Show that Jack can satisfy his goal by asking Fred to pick up block A.

We must here assume that branches in time correspond only to Jack's actions, not to Fred's. Therefore, feasibility for blocks-world actions should be defined in domain-specific axioms, not in terms of branching time. For example, the feasibility of a pickup would be defined in the following rule:

true_in(S,feasible(pickup)) $\Leftrightarrow$
true_in(S,clear(hand)) $\wedge$ $\exists_X$ true_in(S,under_hand(X))

An alternative approach would be to distinguish between what Fred could do in given situation (any physically feasible action) and what he would do (what Jack asks). See Chapter 5, exercise 7.

Assumptions:

i. Starting situation.
true_in(s0, clear(hand)) $\wedge$ true_in(s0, under_hand(blocka)) $\wedge$
[$\forall_X$ true_in(s0, beneath(blockb, X)) $\Leftrightarrow$ X =blocka].

ii. (Definition) A hand command is "pickup" or "putdown" or "move(L)."
hand_command(E) $\Leftrightarrow$
[E = $\prec$pickup$\succ$ $\vee$ E=$\prec$putdown$\succ$ $\vee$ $\exists_L E$=$\prec$move($\downarrow L\downarrow$)$\succ$].

iii. All the hand's activities are deliberate activities of Fred's.
hand_command(E) $\Rightarrow$
[token_of(K,denotation(E)) $\Leftrightarrow$ token_of(K,deliberate(fred,E))].

iv. Fred executes a hand command if Jack asks him to and it is physically possible to do so.

[hand_command(P) $\wedge$ E=denotation(P)] $\Rightarrow$
[occurs(I, E) $\Leftrightarrow$
$\exists_{I1}$ meet($I1, I$) $\wedge$ occur($I1$,illoc(jack, fred, imperative, P)) $\wedge$
can_do(fred,P,start(I))].

v. Fred and Jack are permanently acquainted.
true_in(S,acquainted(fred, jack)).

vi. The sentence "Please execute a pickup" is under all circumstances a request for a pickup.
pronounctation(P,$\prec$Please_execute_a_pickup$\succ$, english) $\Rightarrow$
mode_of(P, K)=imperative $\wedge$ meaning_of(P, K)=pickup.

vii. Jack can always speak any string.
$\forall_{S,P}$ true_in(S, feasible(do(jack,speak(P)))).

viii. If Jack speaks a string, then it will be common knowledge between him and Fred that he has spoken the string. (Note: This assumption is a quick and dirty device to get around specifying the conditions under which a speech of Jack's is known by Fred, and the frame axioms associated with those conditions.)

occur(I,do(jack,speak(P))) $\Rightarrow$
common_know($\prec${jack,fred}$\succ$,
$\prec$occur(@I@,do(jack,speak(!P!)))$\succ$, end(I)).

ix. If Jack now says "Please execute a pickup," then the only illocutionary act that he will complete at any time during his speech is to request a pickup.
[pronunciation(P, $\prec$Please_execute_a_pickup$\succ$, english) $\wedge$
token_of(K,do(jack(speak(P)))) $\wedge$ start(time_of(K)) = s0 $\wedge$
token_of($K1$,illoc(jack, $A2, M, R$)) $\wedge$
end(time_of($K1$)) $\in$ time_of(K)] $\Rightarrow$
$K1 = K$.

x. In s0, Fred has completed any blocks-world action that he started previously.
[E=pickup $\wedge$ E=putdown $\wedge$
E=move(L) $\wedge$ occur(I, E) $\wedge$ start(I) < s0] $\Rightarrow$
end(I) $\leq$ s0.

xi. Jack knows about all blocks-world events.
[hand_command(E) $\wedge$ occurs(I,denotation(E))] $\Rightarrow$
know(jack, $\prec$occur(@I@, $\downarrow E\downarrow$)$\succ$, end(I)).

xii. Fred and Jack have common knowledge of all the blocks-world axioms plus asssumptions i through xi, above.

To prove: Jack can accomplish block B being clear.
can_achieve(jack, clear(blockb), s0).

Proof: Let plan0 be the plan,

$\prec$sequence(do(jack, speak($\prec$Please_execute_a_pickup$\succ$)),
wait_while(pickup))$\succ$.

By the definition of "can_achieve" (axioms KPS.8, KPS.7), we can prove the desired result "can_achieve(jack, clear(blockb), s0)" by showing that Jack knows that plan0 is feasible, that it leads to clear(blockb), and that its knowledge preconditions are satisfied. Since Jack knows all the axioms (assumption xii and CK.1) and we assume consequential closure on knowledge (axiom KNOW.1), it suffices to show that it follows from the axioms that plan0 is a valid plan to accomplish clear(blockb) and that all the knowledge preconditions are met.

From PL1.a it is easily shown that a plan consisting of two actions in sequence is feasible just if the first step is feasible in the starting situation, and the second step is feasible at the end of the first step. In our case, the first step is "do(jack(speak($\prec$Please_execute_a_pickup$\succ$)))," which, by assumption vii, is always feasible. Given the frame axioms on the blocks world, the state of the blocks changes only if a pickup, putdown, or move is executed. Assumption iii states that such events are actions of Fred's. Assumption iv states that Fred will perform such an action only if Jack asks him to. Assumption ix states that Jack does not complete any such request between s0 and the completion of speaking "Please execute a pickup." Assumption x states that in s0, Fred is not in the middle of an action. Therefore, it follows that Fred does not perform any hand actions until the completion of Jack's speaking. Therefore, using the frame axioms of the blocks world, the blocks remain as they were at the start throughout Jack's speech.

Assumptions vi, viii, and xii guarantee that the conditions of IL.3 are satisfied, and that Jack's speech is a request to Fred for a pickup.

By assumption iv, Fred will execute the requested pickup if he can do so; that is, (by definition of "can_do"), if he knows that it is feasible and that its knowledge preconditions are satisfied. Since "pickup" is a constant, the knowledge preconditions are always satisfied. By the axioms of the blocks world, the pickup is currently feasible. Moreover, since Fred knew the state of the blocks at the start (assumption xii), and he knows that he has not committed any deliberate acts (KPG.2), applying assumptions iii and xi, KNOW.1, and the blocks-world frame axioms he knows that the blocks are in the starting position and that the pickup is feasible. Since he knows that the pickup is feasible and that its knowledge preconditions are satisfied, he can execute the pickup; hence, by assumption iv, he will. By the blocks-world axioms, this will result in block B being clear.

Thus, the plan plan0 is feasible and results in block B being clear. By axiom KPS.5, the knowledge preconditions of plan0 are that the knowledge preconditions of the speech are satisfied in the initial situation, and that Jack know that the knowledge preconditions of the "wait" are satisfied in the middle situation. But the speech involves only a constant argument, and so has no knowledge preconditions. By axiom KPS.4, the knowledge preconditions of the action "wait while the pickup takes place" are that Jack is sure that the pickup will take place, and that he will know when it has taken place. Jack can determine that the pickup will take place from the argument above, since Jack knows all the premises of that argument. Assumption xi says that Jack will know when the pickup does take place. Thus, the knowledge preconditions are satisfied.

Finally, again, Jack knows all the premises of the above argument (assumption xi) so, by axiom KNOW.1, he knows that plan0 is feasible, that it accomplishes the goal, and that its knowledge precondition are satisfied. Thus, Jack can achieve block B being clear. Q.E.D.

10.4 Ethics

The responsible builder of an autonomous robot will design its planning module so that it perfers doing right to doing wrong. Further, though the prudent robot will not assume that all other agents are as moral as itself, in order to avoid paranoia it must assume that most other agents rarely break important prohibitions.

A crude representation of ethical valuation can be developed along the following lines: We will consider ethical qualities as characteristics of an action type E in a situation S. We introduce the predicates

Table 10.4 Axioms for Ethical Predicates

ETH.1.	If E is obligatory in S, then it is permitted. Equivalently, it is never obligatory both to do E and not to do it. obligatory$(E, S) \Rightarrow$ permitted(E, S).
ETH.2.	It is always possible to avoid all prohibited actions. $\forall_S \exists_C \forall_{I,E}$ [$I \subset C \wedge$ start(I)=$S \wedge$ occurs(I, E)] $\Rightarrow$ permitted(E, S).
ETH.3	Obligatory actions are better than prohibited ones. [obligatory$(EO, S) \wedge \neg$permitted(EP, S)] $\Rightarrow$ goodness$(EO, S) >$ goodness(EP, S).
ETH.4	Actions can be assumed to be permitted, unless there is reason to believe them prohibited. plausible(true,(permitted(E, S)).
ETH.5	Prohibited actions are generally avoided. plausible($\neg$permitted(E, S), $\forall_I$ [$I \subset$real_chronicle $\wedge$ start(I)=S] $\Rightarrow$ $\neg$occurs(I, E)).

"permitted(E, S)" and "obligatory(E, S)". (Our use of first-order predicates implies that we assume that the ethical qualities of an action are transparent under substitution of equals.) Obligation and permission are dual relations; it is obligatory to do E in S just if it is not permitted not to do E. Formally, we can express this dual relation as follows:

obligatory$(E, S) \Leftrightarrow \neg$permitted(nonoccurrence$(E)$,$S$).

Another representation of ethical values uses a measure space of "goodness" and a function "goodness(E, S)" mapping an event type E and a situation S to a measure of goodness.

Table 10.4 shows some plausible, though perhaps optimistic, axioms on obligation, permission, and goodness.

The main body of a theory of ethical values would be a specification of what acts are permitted, obligatory, prohibited, good, or bad. This would involve both specific rules (e.g., "Killing people is prohibited except under very rare circumstances") and general rules (e.g., "If doing

A will leave everyone happier than doing B, then, in general, A is a better action than B.") Formulating such a collection of rules is far beyond the scope of this book.

10.5 Possession

The following crude theory of possession will handle simple cases. We assume that an object O is owned by at most one agent at a time. The owner of O can change only by virtue of a transfer by its current owner. Use of O by agents other than the current owner are permitted only if the owner permits it. Table 10.5 shows the formal statement of this theory.

This theory is completed by a specification of what actions constitute the use of an object and how the owner of an object permits some particular use of it.

10.6 Appendix A: Conceptual Dependency

Conceptual dependency (CD) is one of the oldest and most influential representations of commonsense knowledge. It focuses on the representation of primitive human actions for the purpose of representing narrative text. Since it deals with issues spanning a number of our chapters, we have postponed discussing it to this final appendix.

CD was developed by Roger Schank and his associates and students over a number of years [Schank 1969, 1975; Schank and Abelson 1977] during which time it underwent continual development. It was not defined within a formal setting. Therefore, our account below, which integrates CD with the theories we have developed in this book, is to some extent a *post hoc* reconstruction. (Our account closely follows the formalization of CD in [Charniak and McDermott 1985].)

CD is primarily concerned with the characterization of primitive human actions. The basic function is the "action($ACTOR, ACTION, OBJECT, SOURCE, DESTINATION$)," which maps its arguments to an event type. $ACTOR$ is the actor of the event. $ACTION$ is one of eleven[5] constants denoting different types of primitive actions. The meanings of $OBJECT$, $SOURCE$, and $DESTINATION$ vary with particular action types. The primitive acts and the meanings of these arguments are shown in Table 10.6.

[5]The exact number and list varied over time.

Table 10.5 Axioms of Possession

Nonlogical symbols:
owner_of(O): Function. The fluent of the owner of O over time.
transfer(O, A): Function. The action of transferring possession of O to agent A.
use_of(E, A, O): Event type E constitutes a use of object O by agent A.
permits(A, E): State type. Agent A permits event type E.

POS.1. Causal axiom of transference: If the owner of O transfers O to $A2$, then $A2$ becomes the owner.
occurs(I,do(value_in(start(I), owner_of(O)), transfer($O, A2$))) $\Rightarrow$
$A2$=value_in(end(I), owner_of(O)).

POS.2. Only the owner can transfer possession.
occurs(I,do($A1$, transfer($O, A2$))) $\Rightarrow$
$A1$=value_in(start(I), owner_of(O)).

POS.3. Frame axiom of transference: Ownership can change only by an act of transference.
[precedes($S1, S2$) $\wedge$
$A1$ = value_in($S1$,owner_of(O)) $\neq$ value_in($S2$, owner_of(O))] $\Rightarrow$
$\exists_{IT,A2}$ intersect(IT,[$S1, S2$]) $\wedge$
occurs(IT, do($A1$,transfer($O, A2$))).

POS.4. Use by others is permissible only if the owner allows it.
[$A2 \neq$ value_in(S,owner_of(O)) $\wedge$ use_of($E, A2, O$)] $\Rightarrow$
[permitted(E, S) $\Leftrightarrow$
true_in(S,permits(value_in(S,owner_of(O)),E))].

The object of an mbuild or mtrans action is a mental construct. The same problems of referential opacity that we have seen in connection with knowledge and belief therefore arise here as well. We will assume a syntactic theory, in which these objects are strings of characters. CD theory distinguishes between the "mental places" of long-term memory and short-term memory. Thus, the action mtrans may be used either for remembering a fact (transferring information from long-term to short-term memory) or for communication (transferring it from one agent to another).

Table 10.6 Primitive Action Types in CD

ptrans — To move *OBJECT* from *SOURCE* to *DESTINATION*.
propel — To exert a force on *OBJECT* in direction *DESTINATION*.
move — To move body part *OBJECT* to *DESTINATION*.
grasp — To grasp *OBJECT* with body part *DESTINATION*.
ingest — To consume *OBJECT* with body orifice *DESTINATION*.
expel — To emit *OBJECT* from body orifice *SOURCE*.
speak — To make sounds *OBJECT*.
mtrans — To transfer information *OBJECT* from *SOURCE* to *DESTINATION*.
mbuild — To mentally construct (imagine, consider, come to believe) *OBJECT*.
attend — To focus sensory organ *OBJECT* on object *DESTINATION*.
atrans — To transfer possession or control of *OBJECT* from *SOURCE* to *DESTINATION*.

Second, CD defines a number of fluents. These are less central to the theory than the primitive acts, and are often generated *ad hoc*. The following are commonly used:

- place(O) — The place of physical object O.
- mloc(M, P) — Mental location P contains mental object M.
- owner_of(O) — Agent owning object O.
- health-val(A) — Physical state of agent A. Characterized in terms of an integral measure space, running from "dead" through "sick" to "perfect-health".
- happiness(A) — Emotional state of agent A. Characterized in terms of an integral measure space, running from "miserable" to "indifferent" to "content" to "ecstatic".

We define the function "change(F, D)" to be the event of fluent F changing in direction D, where D is either "pos" or "neg". For example, "change(happiness(abel),pos)" is the event of Abel becoming happier.

Finally, CD defines a number of causal connectives. The arguments of these connectives fall into three categories:

- Tokens of physical events or actions.

Table 10.7 Causal Connectives in CD

- result(E, S) — Action E brings about state S.
- enable(S, E) — State S makes action E possible.
- disable(S, E) — State S make action E impossible.
- initiate(S, M) — Action or state S initiates mental state M.
- reason(M, E) — Mental state E is the reason for action E.
- instrumental($E1, E2$) — Action $E1$ is instrumental to action $E2$.

- Tokens of physical states or state changes.
- Tokens of mental actions, mental states, or changes to mental state.

Table 10.7 shows the causal connectives used in CD. We abbreviate the above category as "actions," the second as "states," and the third as "mental states."

Using these primitives, augmented with primitives defined earlier or defined *ad hoc*, we can now represent some simple narratives. Table 10.8 shows the CD representation of the following story:

> Marie saw Jessica pushing a toy wagon. Marie wanted to have the wagon. Marie took the wagon from Jessica. Jessica was upset and started to cry.

10.7 References

Common knowledge: Common knowledge has been studied in a variety of fields, including psychology, economics, game theory, philosophy, and computer science, particularly the theory of distributed systems. [Halpern and Moses 1984] is a particularly important and readable paper. Among other results, it shows that a distributed system cannot attain common knowledge of a fact in a series of communications unless it can attain common knowledge of the fact in a single communication. It also discusses several weakenings of the concept

Table 10.8 CD Representation of a Story

Events and states:

EV1: Jessica pushes a wagon.
token_of(ev1, action(jessica, propel, wagon1, ⊥, ⊥)).

EV2: Marie looks at Jessica push the wagon.
token_of(ev2, action(marie, attend, eyes(marie), ⊥, ev1)).

MS3: Marie knows that Jessica pushes the wagon.
token_of(ms3, mloc(mind_of(marie), ev1)).

MS4: Marie wants to have the wagon.
token_of(ms4, goal(marie, eql(owner_of(wagon1), marie))).

EV5: Marie takes the wagon from Jessica.
token_of(ev5, action(marie, atrans, wagon1, jessica, marie)).

ST6: Marie has the wagon.
token_of(st6, eql(owner_of(wagon1), marie)).

MS7: Jessica is upset.
token_of(ms7, change(happiness(jessica), neg)).

EV8: Jessica cries.
token_of(ev1, action(jessica, expel, tears, eyes(jessica), ⊥)).

Causal connections:

initiate(ev2, ms3).
initiate(ms3, ms4).
reason(ms4, ev5).
result(ev5, st6).
initiate(st6, ms7).
reason(ms7, ev8).

Actual occurrence of these events in this order:

[time_of(ev1) ∪ time_of(ev2) ∪ time_of(ms3) ∪ time_of(ms4) ∪
time_of(ev5) ∪ time_of(st6) ∪ time_of(ms7) ∪ time_of(ev8)] ⊂
real_chronicle.

contains(time_of(ev1), time_of(ev2)).
start(time_of(ev2)) ≤ start(time_of(ms3)) ≤ start(time_of(ms4)).
end(time_of(ev1)) ≤ start(time_of(ev5)).
start(time_of(ms4)) ≤ start(time_of(ev5)).
meet(time_of(ev5), time_of(st6)).
start(time_of(st6)) ≤ start(time_of(ms7)).
contains(time_of(ms7), time_of(ev8)).

that are more easily attainable. [Vardi 1988] contains a number of other papers on the subject. Common knowledge has not been extensively studied in the context of commonsense reasoning. As far as I know, the observation that the agent argument of common knowledge is opaque is original here.

Multiagent planning: [Bond and Gasser 1988] contains a number of important papers on multiagent planning. See particularly [Smith and Davis 1981; Georgeff 1983, 1984, 1986; Rosenschein 1982 and Stuart 1985] on cooperative planning, and [Morgenstern 1987] on plans of influence. The use of a theory of plan interaction in narrative understanding is studied in [Wilensky 1978, 1980] and [Dyer 1985].

Communication: High-level AI theories of communication have been extensively studied from many points of view and different degrees of formality. The representation discussed here derives largely from [Morgenstern 1988]. Other important works include [Appelt 1982; Perrault and Allen 1980; Cohen and Levesque 1987; Cohen and Pollack 1987], and [Litman 1985]. The classic philosophical studies of speech acts are [Austin 1961] and [Searle 1969]. See also [Wittgenstein 1958] to get an idea of what a complete commonsense theory of communication and language would have to include.

Ethics: [Sanders 1989] discusses a representation for ethical valuations, possession, and emotions. The formal treatment of ethical values is called *deontic* logic. Deontic logic has been applied to legal reasoning by McCarty [1983]. Philosophical studies of deontic logic include [Von Wright 1968; Prior 1967].

Possession: Many representational systems for narratives (e.g., [Schank 1975]) have used a theory of possession essentially equivalent to the simple one here. Theories that deal with certain aspects of property in greater detail have been developed for use in legal reasoning: see, for example, [McCarty and Sridharan 1980, 1981].

Conceptual dependency: Conceptual dependency was first introduced in [Schank 1969]; this version of the theory had the basic form of actions, with actor, object, source, and destination, but had only one action primitive: "trans." The standard version of CD is presented in [Schank 1975] and [Schank and Abelson 1977].

10.8 Exercises

(Starred problems are more difficult.)

1. * The following is known as the "cheating husbands" problem.

There was once a community with a distinctive system of conventions for dealing with marital infidelity.

(a) Every woman knew of every man in the community, except her own husband, whether or not he was faithful to his wife.
(b) A woman who knew that her husband was unfaithful was obliged to shoot him at midnight as soon as possible when she found out. If a woman shot her husband, the fact was immediately common knowledge.
(c) All women had common knowledge of rules a and b. All women were perfect logical reasoners.

At the time in question, there were k adulterous husbands. One day, at a meeting of the entire community, one of the husbands made the following public statement, witnessed in common by the entire community: "It must be admitted that at least one of us husbands has committed adultery."

k nights later, every adulterous husband was shot by his wife.

Provide a formal justification for the actions of the women.

2. Express the axioms of common knowledge CK.1–CK.5, using the accessibility relation ck_acc.

3. Justify the following inference: Assume (i) that A declares ϕ to B; (ii) that A and B have common knowledge that A is speaking sincerely; and (iii) that A and B have common knowledge that A knows whether ϕ is true. Infer that A and B have common knowledge of ϕ when the communication is complete.

4. Express the following statements using the primitives of Section 10.3.

(a) Mr. Martin said to Mrs. Barrows "I'm sitting in the catbird seat."
(b) Mr. Martin told Mr. Barrows that he had taken cocaine.
(c) Mrs. Barrows told Mr. Martin to leave her apartment.
(d) Mrs. Barrows asked Mr. Martin whether he was drunk.
(e) Mr. Fitweiler told Mr. Martin that Mrs. Barrows had told him (Fitweiler) that Martin had taken cocaine.

5. * Validate the following plan of influence [Morgenstern 1988].

John wishes to call Mary, but doesn't know her phone number. However, he does know that Harry knows her number, and that Harry will tell him the number if asked. He therefore constructs the following plan: Ask Harry for Mary's number; wait for Harry to answer the question; dial the number that Harry states.