Chapter 8

Minds

It was in none other than the black, memorable year 1929 that the indefatigable Professor Walter B. Pitkin rose up with the announcement that 'for the first time in the career of mankind, happiness is coming within the reach of millions of people.' Happy living, he confidently asserted, could be attained by at least six or seven people out of every ten, but he figured that not more than one in a thousand was actually attaining it. However, all the external conditions required for happy living were present, he said, just waiting to be used. The only obstacle was a psychological one. Figuring on a basis of 130,000,000 population in America and reducing the Professor's estimates to round numbers, we find that in 1929 only 130,000 people were happy, but that between 78,000,000 and 91,000,000 could have been happy, leaving only 52,000,000, at the outside, doomed to discontent. The trouble with all the unhappy ones (except the 52,000,000) was that they didn't Know Themselves, they didn't understand the Science of Happiness, they had no Technique of Thinking. Professor Pitkin wrote a book on the subject; he is, in fact, always writing a book on the subject. So are a number of other people. I have devoted myself to a careful study of as many of these books as a man of my unsteady eyesight and wandering attention could be expected to encompass. And I decided to write a series of articles of my own on the subject, examining what the Success Experts have to say and offering some ideas of my own, the basic one of which is, I think, that man will be better off if he quits monkeying with his mind and just lets it alone.

James Thurber, *Let Your Mind Alone!*

It is to the advantage of a thinking creature to be aware of thought and to be able to reason about it. If it has to interact with other thinking creatures, then reasoning about their mental processes is often necessary to understand and predict their behavior. Even if the creature is alone, stranded on a desert island, the ability to reason about its own mental life will be valuable. A Robinson Crusoe robot will need to make plans involving the gaining and maintaining of knowledge, such as "If I want to learn my way around the island, I should go to the top of the big hill," or "If I want to avoid being eaten, I should keep a close watch."

A commonsense mental theory is in many ways harder to formulate than a physical theory. We have no "correct" theory to draw on. The fundamental natures of basic psychological phenomena such as thinking, perceiving, reasoning, and believing are very little understood, despite the efforts of psychology, philosophy, neurophysiology, and AI. Natural language provides a vocabulary that is rich but vague and ambiguous. Our intuitions are often strong, but they are hard to systematize. Moreover, the mental life and its relation to behavior are notoriously lawless. We can almost never make certain predictions about an individual. Even if a prediction is in practice essentially certain, it often seems intuitively that freedom of choice rules out absolute certainty. Whatever rule we put forward, no matter how qualified — say, "If a man is hungry, and there is food set before him ready to be eaten, and he is able to eat it immediately, and he plans to eat it immediately, and he has no reason not to eat it immediately, and he has nothing else on his mind, and he is aware of all this, then he will eat it" — it can be objected that, though a likely conclusion, it is not certain, since he can always act capriciously and not eat it.

Our commonsense understanding of the life of the mind is rich and complex. Consider the following passage from "Ali Baba" in the *Arabian Nights:*

> Cassim rose early in the morning and set out with ten mules laden with great chests, which he planned to fill. He followed the road which Ali Baba had told him. When he came to the door, he pronounced the words "Open Sesame" and it opened. When he was in, it shut again. In examining the cave, he was astonished to find much more riches than he had supposed from Ali Baba's story. He was so fond of riches that he could have spent the whole day in feasting his eyes with so much treasure, if the thought that he came to carry some away with him had not hindered him. He laid as many bags of gold as he could carry away by the entrance. When at last he came to open the door, his thoughts were so full of the great riches he should

> possess that he could not think of the necessary word. Instead of "Open Sesame," he said, "Open Barley," and was very much amazed to find that the door did not open, but remained fast shut. He named several sorts of grain — all but the right one — and the door would not open.
>
> Cassim had never expected such an accident. He was so frightened at the danger he was in that the more he endeavored to remember the word "Sesame," the more his memory failed. He had as much forgotten it as if he had never heard it in his life. He threw down the bags with which he had laden himself, and walked hastily up and down the cave without the least attention to the riches that were around him.

The passage is straightforward, with no surprising or deep psychological insights, and with no mention of the really mysterious aspects of human minds, such as dreaming, consciousness, or intuition. Nonetheless, it presumes a complex theory of mind. Understanding this episode requires knowledge about belief, perception, memory, failures of memory, character traits, goals, plan formation and execution, emotions, and the interactions between all of these. Only a small part of this knowledge has to date been incorporated in formal theories. We do not, at this time, know how to represent most of the knowledge involved in this passage.

One way to avoid some of these problems and complexities is to take our thinkers to be AI programs, or, rather, idealized models of AI programs, rather than people or animals. Here, we can construct a precise underlying theory, we can have definite laws, and issues of free will do not trouble us. We will adopt this device from time to time in our discussion. However, it has the obvious danger of leading us to a consistent dreamworld, in which our theories will be good to describe the behavior of robots constructed according to those very theories, and nothing else.

Our theory of mind will be constructed along the following lines. Certain physical objects — namely, living animals of certain species, including humans, and, possibly, autonomous intelligent robots — are *agents* who have a mental life. We characterize the mental life of an agent in terms of the variations of mental states over time and the occurrence of mental events.

There are many different types of mental states. An agent may hold a proposition to be more or less certainly true; he may be sure of a proposition, believe it, be unsure of it, doubt it, or be sure that it is false. He may hold a proposition to be more or less desirable; he may hope for it, be indifferent to it, or fear it. He may have intentions

to perform certain acts; his deliberate actions, mental and physical, are manifestations of those intentions. He may have emotions — love, hate, anger, enjoyment, pity, and so on — that are directed toward certain things and experiences. He may also have undirected emotions, such as pure happiness or sadness without a particular object.

Mental events include perceiving, in which the agent gains information about the external world; reasoning, in which he combines his existing beliefs to form a new belief; deciding, in which he adopts an intention; remembering; forgetting; creating goals; abandoning goals; and changing emotional state. Some of these mental events are deliberate results of previous intentions.

Some of the features of mental lives differ from one agent to another, and from one period of an agent's life to another. These features are categorized in terms of psychological traits; we say that an agent is affectionate, or greedy, or silly, or a talented writer.

Models such as these are called "folk psychological" models. Though they are fairly natural, every part of them has been challenged. It has been argued that "mental life" is just a category one imposes on things in order to generate certain kinds of explanations, and that therefore it could sometimes be correct to ascribe mental states to an entity of a sort that is not generally viewed as a single agent, such as a thermostat or the nation of France [Dennett 1978, chap.1]. It has been argued, conversely, that having beliefs is inextricably bound up with the use of language, and can only be correctly ascribed to creatures that communicate in language [Davidson 1975]. Behaviorists (e.g., [Skinner 1971]) argue that mental states and events are illegitimate philosophical constructs, and that psychology must be couched in terms of stimuli, responses, and drives. Some dualist philosophers, such as Leibniz, have claimed that the physical world does not actually affect the mental world. The connection between mental states and propositions is the matter of intense debate, both as to its nature and its existence. Regardless of the validity of these objections and alternative approaches, however, the folk psychological model seems to be the most suitable for the formal expression of commonsense knowledge of psychology, and we will focus almost exclusively on this model. (An alternative model from [Rosenschein and Kaelbling 1986] will be discussed briefly in Section 8.5)[1]

We will study only limited parts of the model discussed above. This chapter is concerned only with the mental states of knowledge and

[1]It may be consoling to observe that some of the worst philosophical knots in this domain do not have to be addressed in a commonsense theory. In particular, it seems that we can ignore the "mind-body" problem of how mental states relate to states of the brain.

belief, and the events that affect them. In Sections 1 through 5 of this chapter, we present an idealized model of knowledge and believing at a single moment of time, without considering temporal change. In Section 6, we extend this model to include change in state. In Section 7, we present a representation for perceptions. In Section 8, we consider the consequences of dropping the idealizations of our model. Chapter 9 deals with goals, intentions, decisions, and actions. We will not consider emotions or psychological traits in this book; see the Reference section in this chapter for citations.

8.1 Propositional Attitudes

Many important mental states are *propositional attitudes,* relations between the agent and various propositions about the world. For example, Anne sees that it is raining, she hopes that the sun will come out, she fears that the cellar will flood, she believes that her husband knows that the attic window is open. Here, "sees," "hopes," "fears," and "believes" are types of attitudes; "It is raining," "The sun will come out," "The cellar will flood," and "Anne's husband knows that the attic is open," are propositions. As the last example shows, the statement of a propositional attitude may itself be a proposition.

A propositional attitude is a relation between an agent and a proposition: "Anne sees that it is raining," expresses the relation "sees" between Anne's current mental state and the proposition "It is raining." (As mentioned above, we will ignore the temporal aspects of propositional attitudes until Section 8.6.) Statements of propositional attitudes are an opaque context; they are not invariant under substitution of equal terms. For example, "Oedipus believed that Jocasta was the queen" is not equivalent to "Oedipus believed that his mother was the queen," even though Jocasta was his mother.[2] In Chapter 2 we presented three techniques for representing opaque operators with sentential arguments. In modal logics, first-order logic is extended by the introduction of modal operators as logical symbols. In a possible-worlds representation, the sentential operator is replaced by quantification over a particular class of possible worlds. In syntactic representations, the embedded sentence is replaced by a string that expresses that sentence. The sentential operator is an ordinary first-order predicate, which takes a string as an argument.

[2]Barwise and Perry [1982] point out that there is an extensional use of "see," in which one may say, "Oedipus saw his mother talk to Antigone," even if Oedipus did not know the woman was his mother.

8.2 Belief

A basic propositional attitude is to believe that a proposition is true. The conscious behavior of a rational agent is essentially guided by his beliefs and his goals. He acts in a manner that would satisfy his goals if his beliefs were true.[3] In particular, even if an agent's beliefs are false — that is, they do not correspond to reality — still, his deliberate actions will, in the main, correspond to his beliefs, rather than the actual reality. Therefore, an account of an agent's beliefs is critical for understanding him, particularly if these beliefs are false or incomplete in some significant manner. We use a two-place operator, "believe(A, ϕ)," to mean that agent A believes proposition ϕ. In a modal theory, this is a modal operator; in a syntactic theory, it is a two-place predicate. For the time being, we will consider only beliefs of which the agent is quite confident; we will consider uncertain beliefs in Section 8.4.

It is hard to define exactly what constitutes belief in people. We generally infer a fellow human's beliefs from considering his actions, including his speech, by judging what he might reasonably believe, given what he has learned and perceived; and by assuming that he believes what most people believe, or what most people similar to him believe. These general criteria do not amount to any useful formal definition. We therefore take belief in people to be a primitive relation.

It is possible to be much more specific about what an AI program "believes." Indeed, a major theme of this book is that an AI program should be written and conceptualized so that one can say very precisely what the program believes. An AI program of the kind we have been advocating has a knowledge base that encodes all its beliefs, and the meaning of this knowledge base is firmly defined by semantic definitions. Thus, we can determine what an AI program believes just by printing out its knowledge base and interpreting it according to the semantics. We will adopt this as our model of belief, keeping an eye out from time to time to make sure that its consequences are not violently divergent from our feelings about human belief.

There is still an ambiguity, however. Let us say that our knowledge base is written as transparently as possible, as a collection of assertions, and that the program has some inference engine for answering queries from this knowledge base. There are three possible definitions of what the program believes:

[3] We ignore, of course, many kinds of erratic behavior, e.g., meaningless acts, self-destructive acts, and uncontrolled acts.

i. *Explicit belief*: The program believes anything that is explicitly in the knowledge base.

ii. *Derivable belief*: The program believes anything that the inference engine can derive in a retrieval.

iii. *Implicit belief*: The program believes anything that could be inferred in principle via sound deduction from the knowledge base.

Which is the real belief? It would seem that if the purpose of attributing belief is to predict behavior, then derivable belief (ii) would be best. The program will act on the basis of those facts that it can derive for itself. With realistic inference engines, however, derivable belief is very difficult to characterize, since it depends not only on the state of the knowledge base, but also on the circumstances of the query. For example, an inference engine may allocate different amounts of time to queries, depending on circumstances; whether a result can be derived may depend on the time available. Moreover, these circumstances may change during the process of derivation, possibly even as a result of that process. Constructing a plausible model of an inference engine is thus very difficult. (See Section 8.8 for further discussion.) Taking explicit belief as a primitive does not avoid this problem. In almost all cases, some model of inference must be included; otherwise the theory will be impossibly simple-minded. An intelligent agent should be able, under appropriate circumstances, to go from "This is a tiger" and "Tigers are dangerous" to "This is dangerous," or else it hardly deserves the name of intelligent.

The use of implicit belief as a primitive greatly simplifies the problem of characterizing an inference engine. Implicit belief obeys the principle of *consequential closure:* An agent implicitly believes any fact that deductively follows from his belief. In effect, we approximate the real inference engine with an idealized inference engine capable of drawing any deductive inference arbitrarily quickly. The resultant theory is an approximation of rationality that is simple and elegant. It is sometimes unrealistically powerful. An agent cannot fail to grasp any of the implications of his beliefs. In particular, all agents know all mathematical theorems. Worse, an agent cannot have beliefs that are implicitly contradictory without believing everything. Moreover, deductive inference leaves the set of implicit beliefs unchanged; implicit belief therefore cannot be used to model the action of deductive inference. Reasoning about students learning, or about mathematicians researching, therefore requires a model of explicit belief. Conversely, true rationality implies a great deal more than just competence in deduction. We assume that a rational agent will be able to find reason-

able explanations and generalizations, and to make and revise plausible inferences, as well as making logically sound deductions.

Consequential closure is taken as an axiom of most modal logics. As discussed in Section 2.7.1, it is a necessary consequence of a possible-worlds semantics. Thus a modal logic with a possible-worlds semantics can describe only implicit belief; where a more verisimilar theory of belief is needed, a syntactic theory must be used.

8.2.1 Axioms for Belief

We begin our formalization of belief by considering a variety of axioms and inference rules describing the beliefs of an agent at an instant. Table 8.1 shows a number of axiom schemas for belief expressed in a modal language, together with the corresponding axioms on possible worlds. Table 8.2 shows some possible inference rules applying to belief, some deductive and others plausible. (Further on in this chapter, Table 8.4 will show these axioms expressed in terms of possible worlds, and Table 8.5 will show them expressed in syntactic terms.) It is not necessary to use all these axioms and rules together in a theory of belief; rather, Tables 8.1 and 8.2 should be viewed as a smorgasbord, from which one extracts a subset of axioms suitable to a given application.

We can divide the axioms and rules of Tables 8.1 and 8.2 into three general categories. BEL.1, BEL.2, BEL.9, and BEL.14 discuss the closure of belief under logical or plausible inference. BEL.3, BEL.4, BEL.5, BEL.6, BEL.11, and BEL.13 state weak forms of the principle of veridicality. BEL.7, BEL.8, BEL.10, and BEL.12 characterize introspection. We examine each of these categories in turn.

Logical closure: Implicit belief, by definition, is closed under logical implication. This property is expressed in rules BEL.1 (consequential closure), BEL.2 (belief in the axioms), and BEL.9 (general consequential closure). These rules are part of most modal logics of belief, including any modal logic with a possible-world semantics. Consequential closure is not plausible for explicit belief. BEL.9 is a useful shortcut, not an independent rule; any instance of BEL.9 can be proven from BEL.1 and BEL.2.

BEL.14 extends BEL.1 to cover plausible inference. If we know that A believes that ψ is a plausible inference from ϕ, and we know that A believes ϕ, then, in the absence of contradictory evidence, we may plausibly infer that A believes ψ. In Section 10.3.2, we will give an example of an inference that uses this rule.

Table 8.1 Axioms of Belief

In all of the following, ϕ and ψ are metalinguistic symbols, ranging over all sentences in the modal language. A is an object-language variable ranging over agents. (This table omits the analogues of MODAL.1–MODAL.4, MODAL.11, and MODAL.12, which describe the interface between the modal operator and the predicate calculus.)

BEL.1. Consequential closure: Implicit belief is closed under *modus ponens.*

$\forall_A$ (believe$(A, \phi) \wedge$ believe$(A, \phi \Rightarrow \psi)) \Rightarrow$ believe(A, ψ).

BEL.2. Belief in the axioms: An agent believes the axioms of logic and of belief.

If ϕ is a logical axiom or an axiom of belief,
then $\forall_A$ believe(A, ϕ).

BEL.3. Consistency: No one believes a statement and its negation.

$\forall_A \neg$(believe$(A, \phi) \wedge$ believe$(A, \neg\phi))$.

BEL.4. Privileged access: If an agent believes that he believes ϕ, then he does, in fact, believe ϕ.

$\forall_A$ believe$(A$,believe$(A, \phi)) \Rightarrow$ believe(A, ϕ).

BEL.5. Axiom of coherence: If an agent believes that he does not believe ϕ, then he does not believe ϕ.

$\forall_A$ believe$(A$,$\neg$believe$(A, \phi)) \Rightarrow \neg$believe$(A, \phi)$.

BEL.6. Axiom of arrogance: An agent believes that all his beliefs are true.

$\forall_A$ believe$(A$, (believe$(A, \phi) \Rightarrow \phi)$).

BEL.7. Positive introspection: If an agent believes ϕ, then he believes that he believes ϕ.

$\forall_A$ believe$(A, \phi) \Rightarrow$ believe$(A$,believe$(A, \phi))$.

BEL.8. Negative introspection: If an agent does not believe ϕ, then he believes that he does not believe ϕ.

$\forall_A \neg$believe$(A, \phi) \Rightarrow$ believe$(A, \neg$ believe$(A, \phi))$.

Table 8.2 Inference Rules for Implicit Belief

In all the following rules, the notation $\phi \vdash_A \psi$ means that, if agent A finds sentence ϕ in his knowledge base, then he is entitled to conclude ψ.

Deductive Rules

BEL.9. General inference rule of consequential closure: An agent believes any logical consequence of his beliefs.

If $(\phi_1, \phi_2, \ldots \phi_k \vdash \psi)$ monotonically, then the statement
$\forall_A$ [believes$(A, \phi_1) \wedge$ believes$(A, \phi_2) \wedge \ldots \wedge$ believes$(A, \phi_k)] \Rightarrow$ believes(A, ψ)
is true.

BEL.10. Necessitation: An agent who has ϕ in his knowledge base may conclude that he himself believes ϕ.

$\phi \vdash_A$ believe(A, ϕ).

BEL.11. Optimism. An agent may infer ϕ from the fact that he himself believes ϕ.

believe$(A, \phi) \vdash_A \phi$.

Nonmonotonic Inference

BEL.12. Inference of ignorance: If an agent cannot infer ϕ, he may infer that he does not believe ϕ.

$(\not\vdash_A \phi) \vdash_A \neg$believe$(A, \phi)$.

BEL.13. Principle of charity: Any belief of any agent is likely to be true.

plausible(believe(A, ϕ) ϕ).

BEL.14. Consequential closure on plausible inference: If A believes that ψ is a plausible inference from ϕ, and A believes ϕ, then it is plausible to infer that A believes ψ.

plausible(believe(A,plausible$(\phi, \psi)) \wedge$ believe(A, ϕ),
believe(A, ψ)).

Weak veridicality: Unlike many modal operators, belief does not obey the rule of veridicality $O(\phi) \Rightarrow \phi$; beliefs may be false. However, there are a number of weaker statements that are worth considering as axioms on belief:

BEL.3. (Consistency) No one believes a statement and its negation; an agent's beliefs are internally consistent. This sets a lower limit on sanity. This axiom, in its literal reading, is fairly plausible as a statement about explicit belief; an inference engine can easily ensure that ϕ and $\neg\phi$ are not both in a knowledge base simultaneously. In a theory of implicit belief, where the principle of consequential closure holds, it is both highly implausible — it is not possible for an agent to ensure that his beliefs are internally consistent — and utterly necessary — an agent whose beliefs are internally inconsistent implicitly believes any statement at all. This is an axiom in most modal logics.

BEL.4. (Privileged access) An agent's beliefs about his own beliefs are correct; if he believes that he believes ϕ, then he is right. This is a special case of a more general principle of privileged access, that people's beliefs about their own mental states are correct. The principle is much debated in philosophy (see the Reference section at the end of this chapter), and one can think of cases where it seems to be wrong, such as a neurotic who believes he loves his mother while he actually hates her; but in most commonsense situations, it is quite plausible. Axiom BEL.4 characterizes an agent who carries out the inference rule of optimism, BEL.9. It is a strictly weaker consequence of the axiom of arrogance, BEL.6.

BEL.5. (Coherence) If a person believes that he does not believe ϕ, then he does not believe ϕ.

$$\forall_A \text{ believe}(A,\neg\text{believe}(A,\phi)) \Rightarrow \neg\text{believe}(A,\phi).$$

This is the principle of privileged access applied to nonbelief. It is logically equivalent to "No one ever believes, both that ϕ is true and that he doesn't believe it."

$$\forall_A \neg\text{believe}(A,\phi \wedge \neg\text{believe}(A,\phi))$$

This is true in any reasonable model of belief; a person who believed that ϕ was true but that he didn't believe it would be seriously confused. It is a strictly weaker consequence of the axiom of positive introspection BEL.7, together with the axiom of consistency, BEL.3.

BEL.6. (Arrogance) An agent believes that all his beliefs are true. Note the difference between this and the axiom of privileged access, BEL.4. The axiom of privileged access states that if a person believes that he believes a particular statement ϕ, then he does, in fact, believe ϕ. This axiom makes the stronger statement that every person believes of every statement ϕ that, if he believes ϕ, then ϕ is true. Whether this is true in a theory of implicit belief depends rather subtly on exactly what is meant by "implicit."

BEL.11. (Inference rule of optimism) If agent A deduces that he believes ϕ, then he can add ϕ to his knowledge base. This inference rule is not sound; there are many cases where A will believe false things, and, therefore, believe(A, ϕ) will be true and ϕ will be false. However, it is a *safe* rule in the following sense: If this rule ever takes an agent from a true premise to a false conclusion, then he could have gotten to that conclusion in any case. For "believe(M, ϕ)" is only true (by definition) if ϕ can be inferred from the statements in the knowledge base; and if there is some way that ϕ can be inferred, then we could have used that means to infer it. This inference rule essentially represents the agent's trust in his own rationality.

BEL.13. (Principle of charity) Any belief of an agent is likely to be true.[4] (Note that this inference can be performed by agents other than A.) This seems like a very strong claim and one's natural instinct is to refute it by enumerating all the stupid and wrong things that people do believe. This, however, is probably an illusion, due to the salience of wrong beliefs. The neighborhood crank who believes in astrology and UFO's is a fount of colorful false beliefs; one tends to forget that these few errors are greatly overbalanced by tremendous numbers[5] of true beliefs: the belief that his name is Sam Jackson, the belief that he has a bathroom on the second floor, the commonsense axioms in this book, and so on. This principle cannot generally be adopted as a certain inference; in domains where it can, it is more reasonable to talk of "knowledge" than of "belief." (See Section 9.5.) It is, however, quite a strong plausible inference.

The principle of charity is important in a theory of communication. The basis of communication is that, if A tells something to B, B will generally believe it. Why should B believe it? It seems plausible

[4] The principle of charity is discussed in [Wilson 1959], and in [Davidson 1974], among other places. Davidson views it as a necessary truth: If an agent's beliefs are not mostly true, then we have no way of saying that the agent is rational, no way of reasonably ascribing any beliefs to him, and no way to determine the contents of his beliefs.

[5] Not that there is any obvious way of individuating or counting separate beliefs.

that B reasons as follows: Most utterances are sincere; hence A probably believes what he says. Most beliefs are true; hence what A said is probably true.[6] (See Section 10.3.2.)

Introspection: The last group of rules allows an agent to determine his own beliefs by examining the contents of his knowledge base, and allows an external reasoner to predict the results of such introspection on the part of the agent.

BEL.7. (Positive introspection) If an agent uses the rule of necessitation, BEL.10, then he can be characterized by an external observer as obeying the law of positive introspection: If he believes ϕ, then he believes that he believes ϕ, since he can deduce that he believes ϕ via necessitation. This is an optional axiom in modal logic.

BEL.8. (Negative introspection) If an agent does not believe ϕ, then he believes that he does not believe ϕ. This characterizes an agent who reliably uses the inference of ignorance. It is often a useful axiom. We need something of the kind to predict that agents will sometimes answer "I have no idea" to queries, or will realize that they have to go seek information. In general, however, it is implausible, since an agent does not believe a statement only if it does not follow from anything that he knows, and this is uncomputable, in general. The difficulty is more than theoretical. This axiom is hard to use in practical problems, since it is hard to show that an agent does not believe ϕ, except by showing that he believes $\neg\phi$. The point is illustrated by a well-known example of John McCarthy's.

- A. "Is the President sitting down or standing up at this moment?"
- B. "I haven't the faintest notion."
- A. "Think harder."

How is it that B knows that thinking harder won't get him anywhere?

BEL.10. (Necessitation) If an agent finds ϕ as an assertion in his knowledge base, then he is justified in concluding that he believes ϕ. As we discussed in Section 2.7, this inference rule cannot be turned into a material implication; the axiom schema $\phi \Rightarrow$

[6] This argument is flawed, since the class of beliefs uttered is by no means a representative sample of the class of beliefs. There are biases in both directions. On the one hand, utterances generally deal with interesting beliefs, which are more likely to be wrong than uninteresting beliefs. (Sam Jackson is more likely to talk about UFOs than to tell you where his bathroom is.) On the other hand, responsible speakers tend not to utter uncertain beliefs, and certain beliefs tend to be more reliable than uncertain beliefs.

believe(A, ϕ) would mean that A believes all true facts, and is not valid. However, whenever it is applied it will be valid, since it can only be applied when ϕ is part of the knowledge base; that is, when A does, in fact, believe ϕ. In fact, it has the curious strength that, whenever it is applied, the conclusion believe(M, ϕ) will be true whether or not the assumption ϕ is true. Necessitation is taken as an axiom in most systems of modal logic, though not usually with this interpretation.

BEL.12. (Inference of ignorance) If an agent cannot infer a fact ϕ from his knowledge base, then he may infer that he does not believe ϕ. This type of inference is central to the nonmonotonic autoepistemic inference [Moore 1985b]. It is a problematic rule in a number of respects. The antecedent of the inference is in general uncomputable. It makes the logic nonmonotonic, since the presence of one inference depends on the absence of another. It introduces a circularity into the concept of inference, which may result in there being either no consistent and logically closed set of implicit beliefs for the agent, or many different such sets.

Rules BEL.3, BEL.4, BEL.5, BEL.11, and BEL.13 are plausible in a theory of explicit belief. The rest of the axioms are inappropriate, as, in general, they imply the agent believes infinitely many statements.

As an example of the use of these rules consider the following problem: Harry says "Anyone who believes that all Libras are judicious also believe that all Capricorns are promiscuous. But not all Capricorns are promiscuous." We assume that Harry is speaking sincerely, and therefore believes what he is saying. We wish to infer that Harry does not believe that all Libras are judicious. (Our axioms do not justify the conclusion "Harry believes that not all Libras are judicious.")

We can formalize this as follows: Let pl be the proposition, "All Libras are judicious," and let pc be the proposition "All Capricorns are promiscuous." Our starting assumptions are "believe(harry, $\forall_A$ believe(A,pl) $\Rightarrow$ believe(A,pc))" and "believe(harry, $\neg$pc)". We wish to infer "$\neg$believe(harry,pl)." Table 8.3 shows the justification of this inference in a modal logic containing the above axioms. The logic uses an axiomatic proof theory, based on the axioms of predicate calculus, the above modal axioms of belief, tautological inference, and the necessitation inference rule, as in Section 2.7. (Tautological inferences are left implicit in the proof below.)

Table 8.3 Proof Using the Modal Theory of Belief

No.	Step	Justification
1.	believe(harry,$\forall_A$ believe(A,pl) $\Rightarrow$ believe(A,pc))	Given
2.	believe(harry, believe(harry,pl) $\Rightarrow$ believe(harry,pc))	Consequential closure from 1.
3.	believe(harry,pc) $\Rightarrow$ $\neg$believe(harry,$\neg$pc)	Consistency (axiom).
4.	believe(harry, believe(harry,pc) $\Rightarrow$ $\neg$believe(harry,$\neg$pc))	Belief in the axiom 3.
5.	believe(harry, believe(harry,pl)) $\Rightarrow$ $\neg$believe(harry,$\neg$pc))	Consequential closure from 2 and 4.
6.	believe(harry, believe(harry,$\neg$pc)) $\Rightarrow$ $\neg$believe(harry,pl))	Consequential closure from 5.
7.	believe(harry, $\neg$pc)	Given.
8.	believe(harry, believe(harry, $\neg$pc))	Positive introspection from 7.
9.	believe(harry, $\neg$believe(harry,pl))	Consequential closure from 8 and 6.
10.	$\neg$believe(harry,pl)	Coherence from 9.

8.2.2 Possible Worlds

As discussed in Section 2.7.2, it is possible to express propositions with a modal operator in a language of possible worlds. To apply this technique to belief, we introduce possible worlds as primitive entities in our ontology. A possible world is one particular way that the world could be. The real world is denoted by the constant w0. Any atomic proposition that could potentially be believed or disbelieved must be viewed as a Boolean fluent over possible worlds. As with temporal fluents, we will use the predicate "true_in(W, P)" to mean that state P holds in world W. For example, the sentence "true_in(w6,blond(michelle))" means that Michelle is blond in world w6. (We will also use the function "value_in(W, F)" for fluents that are not Boolean.) We express propositions about belief using an accessibility relation between possible worlds, "bel_acc($A, W1, W2$)". This relation, read "World $W2$ is accessible from world $W1$ relative to the beliefs of A," means that $W2$ is consistent in all respects with the beliefs that A holds in $W1$; any fact that A believes in $W1$ is actually true in $W2$. Facts about which A has no beliefs in $W1$ may go either way in $W2$; if A neither believes nor disbelieves ϕ in $W0$ then there

will be accessible worlds in which ϕ is true and accessible worlds in which ϕ is false. We can thus state that A believes a sentence ϕ by stating that the corresponding fluent holds in all possible worlds.

For example, the statement "Ralph believes that Michelle is blond" may be expressed

$\forall_{W1}$ bel_acc(ralph,w0,$W1$) $\Rightarrow$ true_in($W1$,blond(michelle)).

Belief in compound sentences can be expressed by compounding the consequents of such implications. For example, the sentence "Ralph believes that either Michelle or Agnes is blond" can be expressed

$\forall_{W1}$ bel_acc(ralph,w0,$W1$) $\Rightarrow$
[true_in($W1$,blond(michelle)) $\vee$ true_in($W1$,blond(Agnes))].

This should be distinguished from the stronger statement, "Either Ralph believes that Michelle is blond, or he believes that Agnes is blond" which is expressed,

[$\forall_{W1}$ bel_acc(ralph,w0,$W1$) $\Rightarrow$ true_in($W1$,blond(michelle))] $\vee$
[$\forall_{W1}$ bel_acc(ralph,w0,$W1$) $\Rightarrow$ true_in($W1$,blond(agnes))].

The statement "Ralph believes that someone is blond" is expressed in the form

$\forall_{W1}$ bel_acc(ralph,w0,$W1$) $\Rightarrow$ $\exists_X$ true_in($W1$,blond(X)).

Statements of imbedded belief can be expressed by chaining together accessibility relations. For example, the statement "Ralph believes that Michelle believes that Agnes is blond" is equivalent to "If world $W1$ is accessible from w0 relative to Ralph, then in $W1$ Michelle believes that Agnes is blond," which, in turn, is equivalent to "If $W1$ is accessible from w0 relative to Ralph, then, if $W2$ is accessible from $W1$ relative to Michelle, then Agnes is blond in $W2$."

[$\forall_{W1,W2}$ bel_acc(ralph,w0,$W1$) $\wedge$ bel_acc(michelle,$W1$, $W2$)] $\Rightarrow$
true_in($W2$,blond(agnes)).

In this way, any sentence in the modal language of belief can be translated into the language of possible worlds.

Axioms BEL.1, of consequential closure, and BEL.2, that the agent believes all axioms, must hold in all possible-worlds systems. The remaining axioms of belief enumerated in Table 8.1 correspond to constraints on the structure of belief-accessibility relations. For example, axiom BEL.7 of positive introspection, "believe(A, ϕ) $\Rightarrow$ believe(A, believe(A, ϕ))," corresponds to the constraint that the accessibility relation be transitive,

$$\forall_{A,W1,W2,W3}\ [\ \text{bel\_acc}(A, W1, W2) \wedge \text{bel\_acc}(A, W2, W3)\ ] \Rightarrow \text{bel\_acc}(A, W1, W3).$$

It is easily shown that the axiom of positive introspection is true if the accessibility relation is transitive. Suppose that in $W1$, A does not believe that he believes that ϕ. Translated into the language of possible worlds, this means that in $W1$ there is an accessible world $W2$ in which A does not believe ϕ; that is, there is a world $W2$ accessible from $W1$ such that there is a world $W3$ accessible from $W2$ such that ϕ is false in $W3$. If transitivity holds, then $W3$ is accessible from $W1$. Thus, there is a world accessible from $W1$ in which ϕ is false; that is, in $W1$, A does not believe ϕ. We have shown that if A does not believe that he believes ϕ, then he does not believe ϕ, which is just the contrapositive of the axiom of positive introspection.

It is also possible, though much more difficult, to establish the reverse relation between the modal axiom and the constraint on possible worlds: If $\mathcal{T}$ is a modal theory that obeys the axiom of positive introspection, and also the axioms BEL.1, consequential closure, and BEL.2, belief in the axioms, then there is a model of $\mathcal{T}$ in which the accessibility relation is transitive [Kripke 1975].

Table 8.4 shows the translation of all the axioms for belief into constraints on accessibility relations.

8.2.3 Syntactic Formulation

Expressing the axioms of Section 8.2.1 in a language of strings and syntactic operators involves some slightly subtle considerations. In particular, the correct manipulation of quantified variables within strings requires care. For example, suppose we want to express the axiom "For all N and A, if N is the address of A, then A believes that N is the address of A." We will assume that addresses are strings of characters such as ⌜59_Turnover_Place⌝. Keeping in mind that, in a syntactic theory, the object of a proposition must be a quoted string, a naive guess at a representation would be,

i. $\forall_{N,A}$ address$(N, A) \Rightarrow$ believe$(A,$⌜address(N, A)⌝).

From this, and the statement, "Oscar's address is 59 Turnover Place," the conclusion should follow that "Oscar believes that his address is 59 Turnover Place."

ii. believe(oscar, ⌜address(⌜59_Turnover_Place⌝,oscar)⌝).

Table 8.4 Axioms of Belief in Terms of Possible Worlds

BEL.1. (Consequential closure)

True in any possible-worlds semantics.

BEL.2. (Belief in the axioms)

True in any possible-worlds semantics

BEL.3. (Consistency)

$\forall_{A,W0} \exists_{W1}$ bel_acc$(A, W0, W1)$.

BEL.4. (Privileged access)

$\forall_{A,W0,W1}$ bel_acc$(A, W0, W1) \Rightarrow$
$\exists_{W2}$ [bel_acc$(A, W0, W2) \wedge$ bel_acc$(A, W2, W1)$].

BEL.5. (Coherence)

$\forall_{A,W0} \exists_{W1}$ bel_acc$(A, W0, W1) \wedge$
[$\forall_{W2}$ bel_acc$(A, W1, W2) \Rightarrow$ bel_acc$(A, W0, W2)$].

BEL.6. (Arrogance)

$\forall_{A,W0,W1}$ bel_acc$(A, W0, W1) \Rightarrow$ bel_acc$(A, W1, W1)$.

BEL.7. (Positive introspection)

$\forall_{A,W0,W1,W2}$ [bel_acc$(A, W0, W1) \wedge$ bel_acc$(A, W1, W2)$] $\Rightarrow$
bel_acc$(A, W0, W2)$.

BEL.8. (Negative introspection)

$\forall_{A,W0,W1,W2}$ [bel_acc$(A, W0, W1) \wedge$ bel_acc$(A, W0, W2)$] $\Rightarrow$
bel_acc$(A, W1, W2)$.

However what actually follows from rule i is that Oscar believes ≺address(N, A)≻, which is either meaningless, or means that everything is everyone's address. The problem, as discussed in Section 2.8, is that there is no connection between the quantified variables N and A outside the quotation marks, and the substring ≺N≻ and ≺A≻ inside the quotation marks. The quoted string does not use the symbols N or A, just the characters : N and : A.

What we actually want to say is the following: If N is A's address, then A believes a string of the form "address (< N inside string delimiters >, < name of A >)" where the name of the agent and the address are inserted in the proper place. To express this, we use three functions introduced in Section 2.8. The function "name_of(X)" maps any entity X to a constant string that denotes X. Thus, if Oscar is the father of Harriet, the following is true: name_of(oscar) = name_of(father_of(harriet)) = ≺oscar≻. The function "dbl_quote(P)" takes a string P as an argument, and returns a string with an additional level of quotation. For example, dbl_quote(≺oscar≻) = ≺≺oscar≻≻. (For the interpretation of this notation in terms of tuples of characters, see Section 2.8). The function "apply($O, A1 \ldots Ak$)" takes as arguments a string O, which spells out an operator, and strings $A1 \ldots Ak$, which spell out operands; it denotes the string that spells out the application of the operator to the operands.

We can now state the rule i correctly:

i′ $\forall_{A,N}$ address(N, A) ⇒
believe(A, apply(≺address≻, dbl_quote(N), name_of(A))).

Thus, if A is oscar and N is ≺59_Turnover_Place≻, then dbl_quote(N) = ≺≺59_Turnover_Place≻≻ and name_of(A) = ≺oscar≻. The arguments to the apply are thus, ≺address≻, ≺≺59_Turnover_Place≻≻, and ≺oscar≻, and the value of the apply is ≺address(≺59_Turnover_Place≻, oscar)≻, which is what was desired.

We can compress the notation with some syntactic sugar. Note that any string within string delimiters that is not a simple symbol can be rewritten as the application of an operator string to operand strings; thus,

≺1 + 1 = 2≻ = apply(≺ = ≻, ≺1 + 1≻, ≺2≻) =
apply(≺ = ≻, apply(≺ + ≻, ≺1≻, ≺1≻), ≺2≻)

We adopt the following convention: Let T be a string within string delimiters, and let S be a substring of T. If S is surrounded in T by down arrows ↓, then S itself, rather than its quoted form, is made an argument to the apply function. If S is surrounded by at signs

@, then name_of(S) is made an argument of the apply function. If S is surrounded by exclamation points !, then dbl_quote(S) is made an argument of the apply function.[7] Thus, we can write the above rule more concisely as

$$\forall_{A,N}\ \text{address}(N, A) \Rightarrow \text{believe}(A, \prec\text{address}(!N!, @A@)\succ).$$

or, equivalently,

$$\forall_{A,N}\ \text{address}(N, A) \Rightarrow \text{believe}(A, \prec\text{address}(\downarrow\text{dbl\_quote}(N)\downarrow, \downarrow\text{name\_of}(A)\downarrow)\succ)$$

Table 8.5 shows how the axioms and inference rules for belief given in modal form in Tables 8.1 and 8.2 can be rewritten in syntactic form. We assume that every agent knows a name for himself. (The modal axioms implicitly make the corresponding assumption that every agent knows a rigid designator for himself.)

As discussed in Section 2.8.2, syntactic theories allow the construction of self-referential sentences and sentences that deny themselves. It is often possible to use such sentences to show that a small set of natural axioms on the sentential operators is inconsistent. In particular, in a syntactic theory we can construct a sentence of the form "I do not believe this sentence." Using this sentence, we can show that a syntactic theory of belief is inconsistent if it contains the axiom of consequential closure, BEL.1, the axiom of knowledge of the axioms, BEL.2, and the axiom of coherence, BEL.5. (Exercise 8.3).

8.3 Degree of Belief

Beliefs are held with greater and lesser degrees of certainty. We can incorporate degrees of certainty within our model by introducing a two-place modal function, "d_belief(A, ϕ)," which maps an agent A and a proposition ϕ into a degree of belief, which is a quantity. Thus, the sentence

d_belief(john, grey(clyde)) > d_belief(john, elephant(clyde))

[7] Readers familiar with LISP will recognize this as analogous to the quasi-quote macro. As in LISP, the interpretation of anti-quote marks inside several layers of string delimiters is a potential area for ambiguity. On the half-dozen or so occasions that this arises in the text, I will indicate the scoping of the anti-quote marks by an explicit comment. Except in one place, the scoping is always to the innermost string delimiters. Ultimately, one would need systematic conventions to deal with this horrible problem.

Table 8.5 Rules for Belief: Syntactic Form

In these rules, A is an object-level variable ranging over agents, and P and Q are object-level variables ranging over strings that spell out sentences. ϕ is a metalevel variable ranging over sentences, for use in axiom schemas and inference rules. BEL.9–BEL.14 below are inference rules. BEL.2 is an axiom schema. BEL.1 and BEL.3–BEL.8 are simple axioms, single first-order sentences.

BEL.1. (Consequential closure)

$\forall_{A,P,Q}$ (believe$(A, P) \wedge$ believe$(A, \prec\downarrow P\downarrow\Rightarrow\downarrow Q\downarrow\succ)) \Rightarrow$ believe(A, Q).

BEL.2. (Belief in the axioms)

If ϕ is a logical axiom or an axiom of belief, and string P spells out ϕ, then $\forall_A$ believe(A, P) is an axiom.

BEL.3. (Consistency)

$\forall_{A,P} \neg$(believe$(A, P) \wedge$ believe$(A, \prec\neg\downarrow P\downarrow\succ))$.

BEL.4. (Privileged access)

$\forall_A$ believe$(A,\prec$believe$(@A@, !P!)\succ) \Rightarrow$ believe(A, P).

BEL.5. (Coherence)

$\forall_{A,P}$ believe$(A,\prec\neg$believe$(@A@, !P!)\succ) \Rightarrow \neg$believe$(A, P)$.

BEL.6. (Arrogance)

$\forall_{A,P}$ believe$(A, \prec$believe$(@A@, !P!) \Rightarrow \downarrow P\downarrow\succ)$.

BEL.7. (Positive introspection)

$\forall_{A,P}$ believe$(A, P) \Rightarrow$ believe$(A,\prec$believe$(@A@, !P!)\succ)$.

BEL.8. (Negative introspection)

$\forall_{A,P} \neg$believe$(A, P) \Rightarrow$ believe$(A, \prec\neg$ believe$(@A@, !P!)\succ)$.

Table 8.5: Rules for Belief: Syntactic Form (Continued)

BEL.9. (General inference rule of consequential closure)

If ($\phi_1, \phi_2, \ldots \phi_k \vdash \psi$) monotonically, strings $P_1 \ldots P_k$ spell out $\phi_1 \ldots \phi_k$ respectively, and Q spells out ψ, then the statement

$\forall_A$ [believes$(A, P_1) \wedge \ldots \wedge$ believes(A, P_k)] $\Rightarrow$ believes(A, Q).

BEL.10. (Necessitation)

If string P spells out sentence ϕ, then $\phi \vdash_A$ believe(A, P).

BEL.11. (Optimism)

If string P spells out sentence ϕ, then believe$(A, P) \vdash_A \phi$.

BEL.12. (Inference of ignorance.)

If string P spells out sentence ϕ, then $(\not\vdash_A \phi) \vdash_A$ $\neg$believe(A, P).

BEL.13. (Charity)

If string P spells out sentence ϕ, then plausible(believe(A, P), ϕ).

BEL.14. (Consequential closure on plausible inference)
If A believes that Q is a plausible inference from P, and A believes P, then it is plausible to infer that A believes Q.

plausible(believe$(A, \prec$plausible$(\downarrow P \downarrow, \downarrow Q \downarrow) \succ) \wedge$
believe(A, P),
believe(A, Q))

means that John is more sure that Clyde is grey than that he is an elephant. (In a syntactic theory, the second argument would be a string spelling out a sentence.)

Degrees of belief are generally assumed to be governed by a calculus of uncertainty like those of Chapter 3. For example, if we assume that agents follow a probabilistic model in assigning degrees of belief, we can state the following axioms:

DBEL.1. d_belief$(A, \phi) \in [0,1]$.

DBEL.2. If $\vdash \phi$ then d_belief$(A, \phi) = 1$ and d_belief$(A, \neg\phi) = 0$.

DBEL.3. If $\vdash \phi \Leftrightarrow \psi$ then d_belief(A, ϕ) = d_belief(A, ψ).

DBEL.4. If $\vdash \neg(\phi \wedge \psi)$ then
d_belief$(A, \phi \vee \psi)$ = d_belief(A, ϕ) + d_belief(A, ψ).

We can then interpret the unquantified belief predicate "believe (A, ϕ)" as meaning certain belief:

believe$(A, \phi) \Leftrightarrow$ d_belief$(A, \phi) = 1$.

Axioms BEL.1, BEL.2, and BEL.3 on "believe(M, ϕ)" then follow directly from DBEL.1–DBEL.4 above. Note that we are still assuming a perfectly rational agent with consequential closure here; we are merely allowing him to be unsure of certain things.

How to assign values to embedded statements of belief is not altogether clear. The axiom of privileged access suggests, perhaps, that one's beliefs about one's own beliefs should always be certain. If so, we can reasonably adopt the following axioms:

DBEL.5. d_belief$(A,$ d_belief$(A, \phi) = X) \in \{0, 1\}$.
(An agent is always perfectly sure of his own beliefs.)

DBEL.6. d_belief$(A,$ d_belief$(A, \phi) = X) = 1 \Rightarrow$ d_belief$(A, \phi) = X$.
(An agent is always right about his own beliefs. Privileged access.)

DBEL.7. d_belief$(A, \phi) = X \Rightarrow$ d_belief$(A,$ d_belief$(A, \phi) = X) = 1$.
(An agent always knows his own beliefs. Positive introspection.)

8.4 Knowledge

For an agent to succeed in accomplishing his goals reliably, he must have adequate knowledge of the relevant issues. Having incorrect beliefs about the issues is often worse than useless. Therefore, predicting the success or failure of an agent at a task generally requires reasoning about his knowledge. In particular, predicting our own future successes or failures requires reasoning about our own knowledge. If we find that our knowledge is likely to be inadequate, we may desire to increase it before we address the task. (There is a very good case to be made that knowledge is a more important concept than belief.

For instance, the word "know" is much more frequently used than the word "believe," or any of its synonyms.)

Like believing a fact, knowing a fact is a propositional attitude, a relation between an agent and a proposition. Knowledge is a stronger relation than belief. If an agent knows a fact, then the fact is true, and he believes it. For example, the statement "Naomi knows that tigers are fierce" implies that tigers are, indeed, fierce and that Naomi believes that tigers are fierce. Not all true belief, however, is knowledge. To know a fact requires having some good reason for believing it. An agent should try to act only on beliefs of his that are true; his best policy to achieve this is to act on beliefs that he holds for good reasons. If a gambler believes that he will win the Lottery and he does win it, we do not say that he knew he would win, because he had no reason for his belief, and consequently his belief did not justify his action. However, it is very difficult to state what constitutes good reason for a belief. (It is also not certain that all true beliefs with good reason are considered knowledge; [Gettier 1963] brings some counterexamples.) Therefore, we treat "know" and "believe" as separate primitive propositional attitudes, connected by the axioms in Table 8.6 below.

Knowledge is generally taken to be a less fundamental concept than belief. Unlike belief, knowledge is not a "pure" psychological predicate; it depends on the state of the outer world, not just on the mental state of the agent.[8] For that reason, it is belief, rather than knowledge, that is critical for predicting actions. If John wants a Coke, and he believes that there is a Coke in the refrigerator, then he will go to the refrigerator. It does not matter whether his belief is true or not, until the moment when he can find out whether it is true. Similarly, the principle of privileged access, that an agent's beliefs about his own mental states are correct, does not apply to the mental state of knowledge; it is not the case that, if John believes that he knows that there is a Coke in the refrigerator, then he does know that there is a Coke in the refrigerator.

However, there is another viewpoint in which knowledge is more fundamental. Suppose we have before us a system that seems to be intelligent, but in which we cannot find anything that looks like a knowledge base or like a declarative representation of propositions. When it "does things right" — when its actions are appropriate to whatever goals we attribute to it — then it is reasonable to say that it had knowledge about those aspects of the world that made its actions right. For example, if a robot stops at the edge of a cliff, then it seems

[8] This is a more tenuous distinction than it might appear. It is not clear to what extent belief is independent of the external world. See [Dennett 1981], [Marcus 1986] and [Putnam 1975].

Table 8.6 Axioms of Knowledge

KNOW.1.	(Consequential closure) [know$(A, \phi) \wedge$ know$(A, \phi \Rightarrow \psi)$] $\Rightarrow$ know(A, ψ).
KNOW.2.	(Knowledge of axioms) If ϕ is a logical axiom or an axiom of knowledge, then know(A, ϕ) is an axiom.
KNOW.3.	(Veridicality) know$(A, \phi) \Rightarrow \phi$.
KNOW.4.	(Positive introspection) know$(A, \phi) \Rightarrow$ know$(A$,know$(A, \phi))$.
KNOW.5.	(Negative introspection) $\neg$know$(A, \phi) \Rightarrow$ know$(A, \neg$know$(A, \phi))$.
KNOW.6.	(Necessitation) $\phi \vdash_A$ know(A, ϕ)

reasonable to say that it knew that it should not go further. With more evidence, we may be more detailed and say that it knew that there was a cliff there, and it knew that it could not survive going over a cliff. However, the attribution of false beliefs to a robot that goes over a cliff is much harder. Did it think there was no cliff there? Or did it think it could go over cliffs with impunity? Or was it suicidal? Or was it not behaving rationally at all? This kind of consideration (among others) has led to the development of a quite different theory of mental states from our "folk psychological" model.[9]

Reflecting considerations such as these, an alternative definition of knowledge has been put forward by Rosenschein and Kaelbling [1986]. Here, a multistate machine is said to know proposition P when it is in state S, if, whenever the machine is in state S, P is true. Belief in this theory is defined as uncertain knowledge: The machine, in state S, believes P with certainty α if the probability of P, given that the machine is in S, is equal to α. A limitation of this model is that it makes it hard to give a semantics to embedded knowledge, to changing knowledge, or to knowledge of anything other than the current state of the world.

Table 8.6 enumerates some possible axioms governing the knowledge of a single agent in a modal language. These are largely analogous to the axioms of belief that we have presented above in Table 8.1. Table 8.7 enumerates axioms that relate knowledge to belief.

[9] These considerations are also related to those that led Davidson to claim that the principle of charity is a necessary property of belief (Section 8.2.1).

Table 8.7 Axioms Relating Knowledge and Belief

KB.1.	(Knowledge is belief) If A knows ϕ then A believes ϕ. know(A, ϕ) $\Rightarrow$ believe(A, ϕ).
KB.2.	(Positive introspection: A) If A believes ϕ, then A knows that he believes ϕ. believe(A, ϕ) $\Rightarrow$ know(A,believe(A, ϕ)).
KB.3.	(Positive introspection: B) If A believes ϕ then A believes that he knows ϕ. believe(A, ϕ) $\Rightarrow$ believe(A,know(A, ϕ)).
KB.4.	(Negative introspection) If A does not believe ϕ, then A knows that he does not believe ϕ. $\neg$believe(A, ϕ) $\Rightarrow$ know(A,$\neg$believe(A, ϕ)).
KB.5.	(Arrogance) A believes that, if he believes ϕ, then he knows ϕ. believe(A,believe(A, ϕ) $\Rightarrow$ know(A, ϕ)).

Axioms KNOW.1 and KNOW.2 establish that knowledge is closed under deductive implicature. These axioms thus describe implicit knowledge; applied to explicit or derivable knowledge, they lead to the same problems as applied to explicit or derivable belief. Axiom KNOW.3 is an intrinsic part of the definition of knowledge. Note that it subsumes the analogues for knowledge of axioms BEL.3–BEL.6 and rules, BEL.11 and BEL.13. Axiom KNOW.4 is analogous to axiom BEL.7 and quite as plausible. By contrast, axiom KNOW.5, asserting that agents have the power of negative introspection on their knowledge, is very strong. For that reason, negative introspection is not generally adopted in theories of knowledge that aim at any degree of psychological verisimilitude, though it can be useful in characterizing

certain closed worlds in which the agent's beliefs may be incomplete but cannot be mistaken. For example, in reasoning about a card game, player A can reason that if player B does not know who holds the queen of spades, then B will know that he does not know it.

Axiom KB.1, that knowledge is belief, is part of the definition of knowledge, as discussed above. Axioms KB.2 and KB.3 are slight strengthenings of axiom BEL.7, of positive introspection on belief; KB.4 is a strengthening of BEL.8, negative introspection on belief; KB.5 is a strengthening of BEL.6, the axiom of arrogance. Note that if we were to define knowledge to be exactly true belief, so that KB.1 was a biconditional, then KB.2 and KB.3 would be equivalent to BEL.7, KB.4 would be equivalent to BEL.8, and KB.5 would be equivalent to KB.6.

We have stated the axioms above as axiom schemas in a modal logic where "know(M, ϕ)" is a modal operator. It is straightforward to convert these to a syntactic notation, in the style of Section 8.2.3. Note that a syntactic theory that contains the axioms of veridicality, consequential closure, and knowing the logical axioms can be shown to be formally inconsistent, using the self-referential sentence "I know that this sentence is false."

It is also possible to translate our modal language of knowledge to a language of possible worlds using a relation "know_acc $(A, W1, W2)$," meaning that world $W2$ is compatible with everything that agent A knows in world $W1$. Axioms KNOW.1 and KNOW.2 hold in any such possible-worlds structure. Axiom KNOW.3 corresponds to the statement that every world is accessible from itself. The translations of axioms KNOW.4 and KNOW.5 are exact analogues of the translations of axioms BEL.7 and BEL.8 given in Table 8.1. By combining the two accessibility relations "bel_acc" and "know_acc," it is possible to express sentences involving both knowledge and belief. For example, axiom KB.1, that if A knows ϕ then A believes ϕ, can be expressed in the axiom "bel_acc$(A, W1, W2) \Rightarrow$ know_acc$(A, W1, W2)$." (If $W2$ is consistent with A's beliefs, then it is consistent with A's knowledge.) The translations of KB.2 and KB.3 are left as exercises.

8.5 Knowing Whether and What

Often, we must express the proposition that another agent has knowledge that we ourselves do not. Since we do not have the knowledge in question, we cannot specify in detail what that knowledge is; we can only give a partial description. Examples:

1. Alfred knows whether Sacramento is the capital of California.
2. Sarah knows what is the capital of California.
3. Charles knows how to get to Sebastian's house.
4. Karen knows how to play the piano.
5. Gil knows a lot about the Bronze Age.

In this section, we will discuss representations for propositions like 1, in which an agent knows whether a fact is true, and 2, in which an agent knows the value of a term. In Section 9.3.2 we will discuss the representation of propositions like 3, where "knowing how" can be thought of as knowing a collection of facts that describe a route to Sebastian's house. Representing propositions like 4, where "knowing how" is not obviously a matter of knowing some collection of sentences, or propositions like 5, which involve the notion of a fact being "about" an entity, is very difficult and is not dealt with in this book. ([Ryle 1949] and [Polanyi 1958] discuss the relation of "knowing how" to "knowing that." [Morgenstern 1988] discusses the formal representation of "knowing how.")

We begin by considering these representational problems in a modal language. We introduce the modal operators "know_whether(A, ϕ)," meaning that agent A knows whether ϕ is true, and "know_val(A, τ)," meaning that A knows the value of term τ. Thus, sentences 1 and 2 could be expressed in the forms

know_whether(alfred, sacramento=capital(california))

know_val(sarah, capital(california))

(Like "know" and "believe," "know_whether" and "know_val" are modal operators that create an opaque context for their second argument. The above sentences are not equivalent to "know_whether(alfred, sacramento=sacramento)" or "know_val (sarah, sacramento)".)

We may now consider how these operators are related to "know". In the case of "know_whether" the relation is obvious and unproblematic: A knows whether ϕ is true if he either knows ϕ or he knows $\neg\phi$.

KW.1. $\forall_A$ know_whether$(A, \phi) \Leftrightarrow$ [know$(A, \phi) \vee$ know$(A, \neg\phi)$].

The relation between "know_val" and "know" is generally taken to be the following: An agent knows the value of a term T if he knows some sentence of the form "$C = T$" where C is a constant.[10] This (not by coincidence) fits perfectly with the conventional reading of sentences containing a modal operator within the scope of a quantifier: If O(ϕ) is a modal operator, then the sentence "$\exists_X$ O($\alpha(X)$)" is taken to be true if (roughly) the sentence "O($\alpha(C)$)" is true for some constant C. (More precisely, $\exists_X$ O($\alpha(X)$) is true if there is an object δ such that, if a new constant symbol C is defined to denote δ, then the sentence "O($\alpha(C)$)" is true. See Section 2.7.1.) We can then state the definition of know_val in the following axiom schema:

KW.2. For any term τ, the sentence "know_val$(A, \tau) \Leftrightarrow \exists_X$ know(A,$X = \tau$)" is an axiom.

Corresponding approaches to expressing sentences involving knowing the value of a term can be formulated in possible worlds and in syntactic approaches. In a language of possible worlds, we say that A knows the value of T if T is the same object in all accessible worlds. For example, if Sarah knows that the capital of California is Sacramento, then in all accessible worlds the capital of California is Sacramento; if Sarah is uncertain whether the capital is Sacramento or Los Angeles, then in some accessible worlds it is Sacramento and in others it is Los Angeles. We thus express the sentence "Sarah knows what the capital of California is" in the formula

$\exists_X \forall_{W1}$ know_acc(sarah,w0,$W1$) $\Rightarrow$
X=value_in($W1$,capital(california)).

In the above formula, "capital(california)" is a fluent over possible worlds.

In a syntactic language, we say that Sarah knows what the capital of California is if she knows some string of the form "< name of the capital of California > =capital(california)." In our notation, we would write this as

$\exists_X$ know(sarah,$\prec$@X@=capital(california)$\succ$).

We can thus define knowing the value of a term as a syntactic relation between an agent and a string that spells out the term.

[10] Or a rigid designator. As discussed in Section 2.7.1, we can identify rigid designators with constant symbols without loss of generality.

KW.2. know_val(A, T) $\Leftrightarrow$ know(A,$\prec$@T@ = $\downarrow T\downarrow\succ$).

Any of these notations will support basic inferences about "knowing what". For example, given that Sacramento is the capital of California, and that Sarah knows what the capital of California is, we can use any of these theories to infer that Sarah knows that Sacramento is the capital of California. (Exercise 7).

Universal quantification outside the scope of a modal operator is interpreted similarly. For example, the sentence "Archie knows all the states of the U.S." can be expressed

(1) $\forall_X$ state(X,us) $\Rightarrow$ know(archie, state(X,us)).

In a possible-worlds semantics, this is expressed

(2) $\forall_X$ true_in(w0,state(X,us)) $\Rightarrow$
[$\forall_{W1}$ know_acc(archie,w0,$W1$) $\Rightarrow$
true_in($W1$,state(X,us))].

In a syntactic theory, it is expressed

(3) $\forall_X$ state(X,us) $\Rightarrow$ know(archie,$\prec$state(@X@,us)$\succ$).

Given any of these facts, and the fact "state(alabama,us)," it is possible to prove that Archie knows that Alabama is a state. Note that the above formulas do not specify that Archie knows that the 50 states are all there are; they will remain true if Archie also believes that Guam is a state. (See Exercise 8.)

In general, the above techniques provide mechanisms for making statements of the form "A knows what the τ is" for one particular meaning of "knowing what." The meaning of "knowing what" will depend on the kind of description chosen to be viewed as a rigid designator in the modal theory, or a name in the syntactic theory. For example if we agree that a street address is a rigid designator for a location, then "Sam knows where Jessica lives," expressed

$\exists_X$ know(sam, lives_at(jessica,X))

means that Sam knows the street address. If we agree that a pair of coordinates in a standard reference frame are a rigid designator for a place, then the same sentence means that Sam knows the coordinates of where Jessica lives. Note that we cannot have street addresses and coordinates both be rigid designators, unless we are willing to posit that whenever a street address is known, a coordinate is also known, and vice versa.

In a syntactic theory, we can get around this by introducing a variety of characterizations of descriptions of objects. For example, if we have functions street_address(X) mapping a place to its street address, and coordinates(X) mapping a place to its coordinates, then we can distinguish between the two types of knowledge in the two formulas,

$\exists_X$ know(sam,$\prec$lives_at(jessica,$\downarrow$street_address(X)$\downarrow$)$\succ$)
$\exists_X$ know(sam,$\prec$lives_at(jessica,$\downarrow$coordinates(X)$\downarrow$)$\succ$)

This flexibility is a major advantage of a syntactic theory over a modal or possible-worlds theory.

8.6 Minds and Time

An agent's beliefs and knowledge deal with time and change over time.

To incorporate time into a modal or syntactic language of knowledge and belief is a straightforward application of the techniques of Chapter 5. We convert knowledge and belief into time-varying states either by adding a situational argument to the operators "know" and "believe" or by defining state functions "knowing(A, ϕ)" and "believing(A, ϕ)." Time can be incorporated in the propositional or string argument of know or believe in any of our previous notations. For example, "At 9:00, Warren knew that he was cold and hungry" can be represented in any of the following forms (among others).

know(warren, cold(warren,s900) $\wedge$
hungry(warren, s900), s900).
(Temporal: Extra argument. Knowledge: Modal)

true_in(s900,knowing(warren,
$\prec$true_in(s900,cold(warren)) $\wedge$
true_in(s900,hungry(warren))$\succ$)).
(Temporal: State type. Knowledge: Syntactic.)

know(warren,cold(warren) $\wedge$ hungry(warren),s900).
(Temporal: Modal. Knowledge: Modal)

Having added this temporal component to knowledge and belief, it is necessary to rewrite the previous axioms and rules BEL.1–BEL.14 and KNOW.1–KNOW.8. For the most part, this is a straightforward adding of a single, universally quantified, situational variable to the

axiom, and using it wherever a situational argument is needed. For example, the modal axiom of positive introspection on belief becomes

$$\forall_{S,A}\ \text{believe}(A, \phi, S) \Rightarrow \text{believe}(A, \text{believe}(A, \phi, S), S)$$

A couple of slightly subtle points may be noted:

- The rules BEL.10, BEL.11, BEL.12, and KNOW.7, where the inference was previously restricted by the agent involved, must be rewritten to be restricted by both the agent and the situation. For example, the rule of necessitation, BEL.10, previously written "$\phi \vdash_A \text{believe}(A, \phi)$," must be rewritten "$\phi \vdash_{A,S} \text{believe}(A, \phi, S)$," meaning that if the agent A finds ϕ in his knowledge base at time S, then he may infer that he believes ϕ at time S.
- In syntactic sentences with imbedded belief states, the internal sentence must use the *name* of the situation, a somewhat problematic concept, particularly since in a continuous model there are uncountably many situations. We take the following view: A name of a situation may contain a real number or a symbol with some real-valued parameter (e.g., a line of a given length). The agent can then coin a name for each situation in turn.

An additional operator that is useful for reasoning about knowledge is that of an agent A knowing the current value of a fluent F in a situation S. We represent this using the operator "know_fluent(A, F, S)". For example, the statement "In s800, Reuben knew the time" is represented "know_fluent(reuben, clock_time, s800)". Under the assumption that an agent always knows a name for a situation, this can be defined as follows:

$$\begin{aligned}&\text{know\_fluent}(A, F, S) \Leftrightarrow \\ &[\text{know\_val}(A,\text{value\_in}(S, F),S) \vee \\ &\text{know\_whether}(A,\text{true\_in}(S, F),S)]\end{aligned}$$

In a possible-worlds semantics, this is expressed by stating that the fluent F has the same value in all accessible worlds.

Table 8.8 shows a number of plausible axioms constraining knowledge and belief over time that suggest themselves. These, like the previous axioms governing knowledge and belief at an instant, are somewhat idealized, particularly axiom BT.1, stating that an agent who knows something never forgets it. We assume further that all agents know these axioms, in accordance with axiom KNOW.2.

Axiom BT.1, in particular, is useful for predicting that the agent will be able to predict his own future knowledge. For example, Table 8.9

Table 8.8 Axioms of Belief and Knowledge Over Time

In all the axiom schemas below, ϕ is a metalevel variable, ranging over "anchored" sentences; that is, sentences without temporal indexicals.

BT.1. An agent who knows ϕ in a situation knows ϕ in all later situations.

$$[\ \text{know}(A, \phi, S1) \wedge \text{precede}(S1, S2)\ ] \Rightarrow \text{know}(A, \phi, S2).$$

BT.2. If A believes ϕ, then he believes that he will always believe ϕ.

$$\text{believe}(A, \phi, S1) \Rightarrow$$
$$\text{believe}(A, \forall_{S2}\ \text{precede}(S1, S2) \Rightarrow \text{believe}(A, \phi, S2), S1).$$

BT.3 If A believes that he will believe ϕ in the future, then he believes ϕ now.

$$\text{believe}(A, \exists_{S2}\ \text{precede}(S1, S2) \wedge \text{believe}(A, \phi, S2), S1) \Rightarrow$$
$$\text{believe}(A, \phi, S1)$$

shows how we can use this to infer that if an agent knows the physics of the blocks world, knows the starting state, and knows that he can trace what is happening, then he can predict that he will know the final state.

In general, however, there are serious problems in formulating a theory in which an agent's future beliefs and knowledge can be predicted. It is difficult to find any reasonable causal or frame axioms on an agent's beliefs. There are also difficulties in constructing a theory that allows an agent to make predictions without assuming that the agent knows of all the events that occur [Morgenstern 1989].

8.6.1 Situations and Possible Worlds

The situation-based theory of time and the possible-worlds semantics for belief and knowledge use two different types of possible worlds. A temporal situation is a snapshot of the world at a given instant; an epistemically possible world is a way in which the world could possibly be. In combining these two logics, it is necessary to connect these two concepts. Moore [1980] has shown that identifying epistemically

Table 8.9 Sample Inference of Agent's Knowledge of the Future

Given:

The above axioms of knowledge.

Daniel knows all the axioms of the blocks world (Table 5.2).

In situation s0, Daniel knows all "beneath" and "place" fluents.

$\forall_{X,Y}$ know_fluent(daniel,beneath(X, Y),s0).

$\forall_X$ know_fluent(daniel,place(X),s0).

Daniel knows that he knows all the blocks.

know(daniel, $\forall_X$ block(X) $\Rightarrow$ know(daniel,block(X),s0), s0).

Daniel knows that he will know whatever events occur when they are done.

know(daniel,$\forall_{I,E}$ occur(I, E) $\Rightarrow$
know(daniel,occur(I, E),end(I)), s0).

Daniel knows that either a pickup, a putdown, or a move occurs in the interval [s0,s1].

know(daniel, occur([s0,s1],pickup) $\vee$ occur([s0,s1],putdown) $\vee$
$\exists_L$ occur([s0,s1], move(L)),
s0).

Infer:

Daniel knows now that in s1 he will still know all beneath relations.

know(daniel,
$\forall_{X,Y}$ know_whether(daniel,beneath(X, Y), s1), s0).

Sketch of the proof: Daniel knows in s0 that in situation s1 he will know what event has occurred in [s0,s1] (Given). By KNOW.1, he knows axiom BT.1, that he will still know everything in s1 that he knows now. In particular, in s0 he knows that in s1 he will still know the positions of all the blocks in s0 and the axioms governing the blocks world. The positions of the blocks in s1 follow logically from their positions in s0, the event in [s0,s1], and the axioms of the blocks world. Hence, applying consequential closure to Daniel's knowledge in s1, it follows that in s1 he will know the positions of the blocks in s1. Since, by KNOW.1, Daniel knows in s0 that his knowledge in s1 obeys consequential closure, it follows, by consequential closure on his knowledge in s0, that he knows in s0 that he will know in s1 where all the blocks are.

possible worlds with temporal situations gives a theory that is elegant and powerful. (The language we construct below is slightly more expressive than Moore's, which represented time using the situation calculus in the narrow sense (Section 5.8).)

We presume a set of parallel chronicles and an accessibility relation between situations. Situation $S2$ is accessible from $S1$ relative to A if $S2$ is consistent with everything that A knows in $S2$. If an agent knows different things in situation s1 than he does in s2, then different worlds will be accessible to him in s1 than in s2.

Knowledge about the past and the future is expressed as statements about the chronicles containing accessible situations. For example, the sentence "Eva knows that Columbus was alive in 1492" is interpreted "In every chronicle containing a situation compatible with Eva's knowledge, Columbus was alive in 1492."

$$\forall_{S1}\ \text{know\_acc(eva,s0,}S1) \Rightarrow$$
$$\exists_{S2}\ \text{precedes}(S2, S1) \wedge \text{true\_in}(S2,\text{alive(columbus)}) \wedge$$
$$\text{value\_in}(S2,\text{clock\_time}) \in \text{year\_1492}.$$

Here, s0 is the current real situation; $S1$ is any situation which, so far as Eva knows, might be the current situation; and $S2$ is some situation in the same chronicle as $S1$ in 1492 when Columbus was alive (Figure 8.1).

Axioms BT.1–BT.3 can be expressed as constraints on the interrelations on temporal precedence and knowledge accessibility. For example, axiom BT.1, that knowledge is never lost, corresponds to the following constraint: If $S1A$ precedes $S2A$ in chronicle A, and $S2B$ is knowledge accessible from $S2A$, then there is a situation $S1B$ that is knowledge-accessible from $S1A$ and that precedes $S2B$ in chronicle B (Figure 8.2).

$$\forall_{S1A,S2A,S2B}\ [\ \text{precedes}(S1A, S2A) \wedge \text{know\_acc}(A, S2A, S2B)\ ] \Rightarrow$$
$$\exists_{S1B}\ \text{precedes}(S1B, S2B) \wedge \text{know\_acc}(A, S1A, S1B).$$

We may justify this formulation as follows: Assume that the above constraint holds. Suppose that A does not know ϕ in $S2A$. Then there is a knowledge-accessible situation $S2B$ in which ϕ is false. If $S1A$ precedes $S2A$ then, by the above constraint, there is a situation $S1B$ that is knowledge accessible from $S1A$ and that precedes $S2B$. Since ϕ is a time-independent sentence, and ϕ is false in $S2B$, ϕ must also be false in $S1B$. Thus, A does not know ϕ in $S1A$. We have thus shown that, if A does not know ϕ at a later time, then he does not know ϕ at an earlier time, which is just the contrapositive of the axiom of memory. The expression of the remaining rules BT.2 and BT.3 in a possible-worlds semantics is left to the reader (Exercise 4b).

8.7 Perceptions

Perceptions are the interface that allows the mind to gain information about the external world; they are the ultimate source of most beliefs. Despite their importance, however, there has been little work to date at developing a commonsense theory of the senses. (The detailed models of the senses provided by vision and other sensory research do not enter into a commonsense understanding.) We will discuss some of the issues involved, and briefly sketch a possible theory.

Consider the following quotation:

> Suddenly, there was the momentary gleam of a light in the direction of the ventilator, which vanished immediately, but was succeeded by a strong smell of burning oil and heated metal. Someone in the next room had lit a dark lantern. I heard a gentle sound of movement, and then all was silent once more, though the smell grew stronger. For half an hour I sat with straining ears. Then suddenly another sound became audible — a very gentle, soothing sound, like that of a small jet of steam escaping continually from a kettle. The instant that we heard it, Holmes sprang from the bed, struck a match, and lashed furiously with his cane at the bell-pull.
>
> "You see it, Watson?" he yelled. "You see it?"
>
> But I saw nothing. At the moment when Holmes struck the light, I heard a low, clear whistle, but the sudden glare flashing into my weary eyes made it impossible for me to tell what it was at which my friend lashed so savagely. I could, however, see that his face was deadly pale and was filled with horror and loathing. (Sir Arthur Conan Doyle, "The Adventure of the Speckled Band," *Adventures of Sherlock Holmes.*)

Understanding this passage involves a rich theory of perception. First, the reader must be able to connect Watson's and Holmes's perceptions to their mental state. Watson infers that a dark lantern has been lit from seeing the light and smelling the oil and the heated metal of the dark lantern. Watson cannot identify the source of the steam-like sound; Holmes, presumably, has identified it. Whatever Holmes has seen is the source of his horror. Second, the reader must connect the perceptions to the physics of the external world. Holmes lights a match in order to be able to see, and Watson sees Holmes by the light he has lit. Third, the reader must know something about the actual sensors; in particular, if one's eyes are used to the dark and are suddenly exposed to bright light, one may be temporarily unable

to see in the direction of the light. An adequate commonsense theory must deal with all these issues.

A representation for statements about perception, which allows the expression of simple rules connecting the physics of the perceivable neighborhood to the knowledge gained through perception, may be developed in a modified possible-worlds semantics on the following lines: We define a *behavior* to be a possible history of the *physical* world over time. We define a *layout* to be a possible instantaneous snapshot of a behavior. Behaviors and layouts are thus analogous to intervals and situations respectively, except that they do not incorporate nonphysical aspects of the world. In particular, agent's beliefs and knowledge are not aspects of a layout. We define a predicate "pc($A, L0, L1$)," read "Layout $L1$ is perceptually compatible with layout $L0$ relative to agent A," meaning that layout $L1$ is consistent with everything that A can perceive in $L1$. We define the function "layout(S)" as giving the physical layout in situation S. The statement that A perceives the value of a fluent F is expressed by asserting that F has the same value in every compatible layout. For example, the statement that in situation s0 Caroline sees that the cat is on the sofa is represented

$$\forall_{L1}\ \text{pc(caroline,layout(s0)},L1) \Rightarrow \text{true\_in}(L1,\text{on(cat15,sofa8)}).$$

(We extend "true_in" and "value_in" to layouts in the obvious way.)

The power of a perceptor are expressed by giving necessary conditions for two layouts to be perceptually compatible; its limits are expressed by giving sufficient conditions. Consider, for example, a robot "r2d2" with sonar. The statement that the robot can always perceive whether or not there is a solid object within distance d0 of it can be represented by stating that two layouts are compatible only if either both have an object within distance d0, or neither does.

$$\begin{aligned}
&\forall_{L1,L2}\ \text{pc(r2d2},L1, L2) \Rightarrow \\
&[[\ \exists_X\ \text{solid}(X) \wedge \\
&\qquad \text{value_in}(L1,\text{distance(place(r2d2),place}(X))) < \text{d0}\ ] \Leftrightarrow \\
&[\ \exists_X\ \text{solid}(X) \wedge \\
&\qquad \text{value_in}(L2,\text{distance(place(r2d2),place}(X))) < \text{d0}\ ]].
\end{aligned}$$

The statement that the robot can never perceive an object more than distance d1 can be expressed by saying that, if two layouts have identical objects within distance d1, then the two are perceptually compatible.

$$\forall_{L1,L2}\ [\ \forall_X\ [\text{value\_in}(L1,\text{distance}(\text{place}(\text{r2d2}),\ \text{place}(X))) < \text{d1} \vee \text{value\_in}(L2,\ \text{distance}(\text{place}(\text{r2d2}),\text{place}(X))) < \text{d1}\ ] \Rightarrow \text{value\_in}(L1,\text{place}(X)) = \text{value\_in}(L2,\text{place}(X))\ ] \Rightarrow \text{pc}(\text{r2d2},L1,L2).$$

Perceptions are connected to knowledge by the rules that an agent who perceives ϕ knows ϕ and knows that he perceives ϕ. These rules are expressed in the following axioms.

PERC.1. $\text{know\_acc}(A,S1,S2) \Rightarrow \text{pc}(A,\text{layout}(S1),\ \text{layout}(S2))$.

PERC.2. $[\ \text{know\_acc}(A,S1,S2) \wedge \text{pc}(A,\text{layout}(S2),\ L3)] \Rightarrow \text{pc}(A,\ \text{layout}(S1),\ L3)$.

Similarly, we represent the perception of an event by defining a compatibility relation "bpc($A, B1, B2$)" on behaviors. Behaviors $B1$ and $B2$ are compatible relative to agent A if, as far as A can see during I in $B1$, the world might be going through $B2$. Using the function "behavior(I)" mapping an interval I to its behavior, we can represent a sentence like "In interval i0, Hector saw the rabbit eat the carrot" in the form,

$$\forall_{B1}\ \text{bpc}(\text{hector},\ \text{behavior}(\text{i0}),\ B1) \Rightarrow \text{occurs}(B1,\text{eat}(\text{rabbit1},\ \text{carrot1})).$$

8.8 Realistic Models of Mind

Once we drop the idealization of complete and error-free reasoning and perception, we enter *terra incognita*. As mentioned at the beginning of this chapter, the actual commonsense theory of mind is very rich, and only very limited and preliminary formal models of this theory have been constructed. The most we can do, at this stage, is to discuss a few of the prominent issues that come up in this theory.

1. A complete commonsense theory of mind must include all mental activities of which we are naturally aware. This does not mean that the commonsense theory of mind need include all of cognitive psychology, because we are only aware of the high-level structure of these activities, not of their fine detail. Thus, a commonsense theory of vision need not contain any account of the mechanisms of vision, because these mechanisms are not known to common sense.

(If they were, vision research would be a lot easier.) What the commonsense theory needs is high-level characterizations; rules such as "Familiar objects can generally be recognized." Note that we could have a complete algorithmic theory of vision without having identified these rules. The two problems are quite separate.

2. Mental activities, including deductive inference, generally take perceptible amounts of time. Therefore, these should be considered a type of event and connected by causal theories.

3. Since beliefs are not closed under logical implication, possible-worlds and standard modal theories are out of the question, and syntactic theories must be used.[11] It is not even reasonable to require that beliefs be logically consistent, since logical consistency is noncomputable. It may be reasonable, however, to demand some level of coherence; to require, for example, that if a person at one time believes both P and $\neg P$, then he will try to resolve this conflict.

4. Mental events are sometimes deliberately planned. One may plan to think about a problem, or to pursue some particular line of thought, or to remember some particular item. Not all mental events can be planned, or one ends in an infinite regress [Haas 1986]

5. To reason about agents that can forget something and then remember it later, we must use a model with at least two different knowledge bases of different functionalities. One knowledge base, short-term memory, contains the beliefs of which the agent is currently aware; the other, long-term memory, contains everything that he has known. Remembering is the event of a proposition in long-term memory coming to short-term memory. Note that it is possible to be aware in short-term memory that one knows a fact in long-term memory but not in short-term memory; for example, Cassim in the Ali Baba story is aware that he knows the password in long-term memory, but he can't remember it. This means that propositions in the two knowledge bases must be able to refer to the knowledge bases. Common sense includes a fair amount of knowledge about the interactions between the two knowledge bases. For example, the Ali Baba story quoted at the beginning of the chapter relies on the knowledge that distraction and terror make remembering difficult. The climax of *The Prince and the Pauper* rests on

[11] The theories proposed in [Levesque 1984] and [Fagin and Halpern 1985], which are not closed under logical implication, combine aspects of modal and syntactic approaches.

the common observation that an event can be remembered if the course of events leading up to it is rehearsed.

6. The varying knowledge and mental powers of different people, or of the same person over different times, can be categorized in general terms within a commonsense theory. "Cassandra knows a lot of differential geometry." "Richard has perfect pitch." "Edmund is very gullible." "Elaine has an excellent memory, but no sense of direction."

7. It is sometimes possible and useful to state a rule of the form "If ϕ were true, then A would know ϕ"; for example, "If Bob had an older brother, then he would know about it." Generally, this rule is used to infer that, since A does not know ϕ, ϕ must be false. Such rules are derived from general knowledge about what A may be presumed to have learned. We know that most siblings get to be known as part of the family circle, and that, in the rare cases where they are not, they will usually be spoken of from time to time. If Bob had been born when his parents were relatively old, and it was known that his parents had many skeletons in the closet that they never discussed, then the inference rule might not apply. The background knowledge that lies behind rules like this has not been much studied.

8.9 References

One could spend several lifetimes reading the philosophy of mind. Most of the central issues in the theory of knowledge can be found in the Platonic dialogue *Theatatus*. I personally have found Bertrand Russell's *Human Knowledge: Its Scope and Limits* [1948] extremely enlightening. [Ryle 1949] raises many interesting and difficult issues in the theory of mind; his approach is totally at variance with the approach taken here. The articles on "Epistemology, History of" and on "Knowledge and Belief" in the *Encylopedia of Philosophy* are good general surveys.

The classic works on the modal theories of knowledge and belief were written by Hintikka [1962, 1969]. This modal theory of knowledge was applied to AI and related to the theory of temporal situations by Moore [1980, 1985a]. Moore [1980] also gives a first-order axiomatization of a scheme to translate between modal representations and possible-worlds representations. [Halpern and Moses 1985] reviews the modal theories of knowledge and belief, and adduces complexity results in the propositional theory. Levesque [1984] examines a weak-

ened modal logic of explicit belief that avoids the assumption of consequential closure; extensions of this approach are studied in [Fagin and Halpern 1985]. [McCarthy et al. 1978] was an early AI work on the modal theory of knowledge. [Halpern 1986] and [Vardi 1988] are collections of research papers on formal theories of knowledge.

Syntactic theories of belief and knowledge were studied by Kaplan and Montague [1960], who showed that these theories lead to paradoxes of self-reference. See also [Thomason 1980] for a discussion of these paradoxes. The model of belief as membership in a knowledge base was put forward in [Moore and Hendrix 1982]. Konolige [1982] and Haas [1983] have used syntactic theories to construct models of limited inference engines that lack consequential closure. Morgenstern [1988] has used a syntactic theory of knowledge with consequential closure to achieve a flexible language for describing knowledge preconditions for plans (see Chapter 9). Konolige, Haas, and Morgenstern each propose a possible solution to the paradoxes of self-reference. The syntax used in Section 8.2.3 follows Morgenstern, with minor notational variants.

The axioms of knowledge and belief discussed in the text are mostly culled from modal logic; see, for example, [Hughes and Cresswell 1968]. The principle of charity is discussed in [Davidson 1974]. [Rosenschein and Kaelbling 1986] presents the alternative model of knowledge and belief discussed in Section 9.5. The axiom of arrogance is presented in [Kaplan and Montague 1960]. [Moore 1985b] discusses the inference of ignorance, and the inference "Since A does not know ϕ, ϕ must be false." The principle of privileged access is discussed in [Kripke 1972]

[Hintikka 1962], [Moore 1980] and [Moore 1985a] discuss quantification in modal theories of knowledge. [Haas 1982] and, particularly, [Morgenstern 1988] discuss the partial specification of propositions in syntactic theories.

The relation between possible-worlds semantics for knowledge and situation semantics for time is studied in [Moore 1980] and [Moore 1985a]. [Morgenstern 1989] discusses the difficulties of integrating theories of knowledge with solutions to the frame problem.

[Gettier 1963] argues that "Justified true belief" is not an adequate definition of knowledge. [Putnam, 75], [Dennett 1981], and [Marcus 1986] bring arguments that suggest that what an agent may be said to believe depends on his relation to the external world, as well as his internal state.

There is little of substance on commonsense theories of perception. Conceptual dependency theory [Schank 1975] included ATTENDing a

sensor as a primitive act, which could be causally connected to mental events. The theory of layouts sketched above is developed in [Davis 1988, 1989].

There are no very detailed unidealized models of mind. [Konolige 1982] gives general schemes for limiting the power of an inference engine. [Haas 1983] shows how inference may be treated as an event; [Haas 1986] extends this to show how it can be treated as a planned action. [Thomason 1987] discusses a number of interesting aspects of belief, and presents a partial model. The use of a two-level theory of memory is an old idea in cognitive psychology. It was used in conceptual dependency theory [Schank 1975] with the primitive act MTRANS to transfer information from one to the other, but then it fell out of interest. [Kube 1985] deals with a number of interesting issues in this theory.

For formal theories of emotion, see [Schank and Abelson 1977]; [Roseman 1979], [Lehnert 1980], [Dyer 1983], and [Sanders 1989].

8.10 Exercises

(Starred problems are more difficult.)

1. Represent the following sentences (i) in a modal language; (ii) in a language of possible worlds; and (iii) in a syntactic language. For this exercise, ignore the temporal component of these sentences. (Also ignore the fact that, in English, "A does not know ϕ" is always interpreted to mean that ϕ is true, despite A's ignorance. For the purposes of all the exercises in this chapter, treat "A does not know ϕ" as meaning no more than "It is false that A knows ϕ.")
   (a) Jack knows that he lives in Hertfordshire.
   (b) Jack believes that Algernon does not know that Jack lives in Hertfordshire.
   (c) Algernon knows where Jack lives.
   (d) Jack does not know that Algernon knows where Jack lives.
2. Represent the following sentences (i) in a modal language; (ii) in a language of possible-worlds; and (iii) in a syntactic language. Be sure to represent the temporal component of these sentences.
   (a) In situation s1, Algernon knew that Jack lived in Hertfordshire.
   (b) In situation s2, Jack did not know that Algernon had already been at his house for an hour.

(c) In situation s3, Jack did not believe that Algernon would remain at his house for a week.

(d) In situation s4, Jack knew that Cecily had not ever believed that Algernon was dead.

3. Show how the following inferences can be carried out (i) in a modal logic; (ii) in a logic of possible worlds; and (iii) in a syntactic logic. (You may ignore the temporal component.)

   (a) Given: Lord Bracknell does not know where Gwendolen is.
   Infer: Lord Bracknell does not know that Gwendolen is in Hertfordshire.

   (b) Given: Algernon knows that Aunt Augusta believes that Bunbury is sick.
   Infer: Aunt Augusta does not know that Bunbury is not sick.

4. *

   (a) For each of the axioms BEL.3, BEL.4, BEL.5, BEL.6, and BEL.8 show that, if the possible-worlds axiom in Table 8.4 holds, then the corresponding modal axiom in Table 8.1 holds.

   (b) Express axioms BT.2 and BT.3 in the language of possible-worlds.

5. * Show that a syntactic theory is inconsistent if axioms BEL.1, BEL.2, and BEL.5 hold, and the sentence "I do not believe this sentence" can be constructed.

6. * Show that the axiom of arrogance (BEL.6) is strictly stronger than the axiom of privileged access (BEL.4) by exhibiting a possible-worlds structure in which the axiom of privileged access holds but the axiom of arrogance does not. (Hint: All you need are two worlds and one atomic formula.)

7. Given that Sacramento is the capital of California and that Sarah knows what the capital of California is, show that Sarah knows that Sacramento is the capital of California (a) in a possible-worlds theory; and (b) in a syntactic theory.

8. * Represent the statement "Archie knows exactly what the states of the U.S. are" in a formal sentence ϕ. Given ϕ and the fact "state(alabama,us)," it should be possible to infer that Archie knows that Alabama is a state; given ϕ and the fact "$\neg$state(guam,us)," it should be possible to infer that Archie knows that Guam is not a state. ϕ should not contain the names of all 50 states. (Hint: Use set theory.)