Chapter 1

A Plea for the Study of Reasoning

Reasoning as Reasoned Change in View

Intending to have Cheerios for breakfast, Mary goes to the cupboard. But she can't find any Cheerios. She decides that Elizabeth must have finished off the Cheerios the day before. So, she settles for Rice Krispies. In the process, Mary has modified her original intentions and beliefs.

This is a very simple case of reasoned change in view, an elementary example of reasoning. It has the following two features. First, not only does Mary's reasoning lead her to add new beliefs to her view, so that she comes to believe that there are no more Cheerios and that Elizabeth ate the last Cheerios yesterday, it also leads her to give up things she had been believing, so that she stops believing that there are Cheerios in the cupboard and that she will have some Cheerios for breakfast. Second, Mary's reasoning changes not only her beliefs but also her plans and intentions. Her reasoning leads her to abandon her intention to have Cheerios and to adopt the new plan of having Rice Krispies. In other words, her reasoning is not only "theoretical," affecting her beliefs, but also "practical," affecting her intentions and plans.

In saying this, I assume that Mary's reasoning can be separated into distinct segments of practical and theoretical reasoning. I make this assumption even though any given segment of reasoning is likely to affect both her beliefs and her intentions, since changes in her beliefs can affect her plans, and changes in her plans can affect her beliefs. When Mary stops believing there are any Cheerios left, she also stops intending to have Cheerios for breakfast. When she forms her intention to have Rice Krispies instead, she also comes to believe that she will be having Rice Krispies for breakfast. But I assume there is a difference between immediate changes that are "part of" a given segment of her reasoning and less immediate changes that are merely further effects of that segment of reasoning.

True, it is not easy to say when a change is "part of" a given segment of reasoning and when it is merely the result of reasoning. For example, it is not immediately obvious whether changes in desires are *ever* part

of reasoning. (I discuss this question briefly in chapter 8.) Nevertheless, in what follows I assume there is a definite difference between immediate changes that are part of a given segment of reasoning and other less immediate changes that are merely further effects of it. And I also assume there is a distinction between theoretical and practical reasoning.

These assumptions suggest a further distinction between two sorts of rules of reasoning, corresponding to two possible phases in reasoning. On the one hand there is often a process of reflection in which one thinks about one's beliefs, plans, desires, etc. and envisions various possibilities in more or less detail. On the other hand there is the actual revising of one's view, which may or may not follow such a reflection. *Maxims of reflection*, as we might call them, say what to consider before revising one's view, for example, that one should consider carefully all the alternatives, with vivid awareness of relevant evidence of possible consequences of contemplated courses of action. On the other hand what we might call *principles of revision* concern the actual changes to be made, the changes that are actually "part of" the reasoned revision, saying such things as that one should make minimal changes in one's view that increase its coherence as much as possible while promising suitable satisfaction of one's ends.

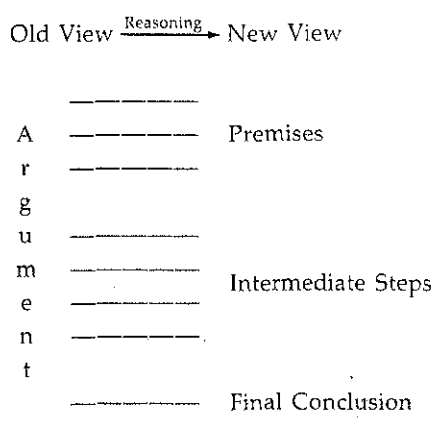
Not all principles of psychological change are principles of revision in this sense, since not all changes are instances of reasoning. For example, it may be that changes in desires are not instances of reasoning, although these changes can occur as a result of reasoning. Even so, there may be general principles governing changes in desires. These would be principles of change that were not principles of revision in the relevant sense.

I don't want to suggest that one ever makes *conscious* use of principles of revision in changing one's view. One can reason without knowing what the relevant principles of revision are and it may well be that reasoning is a relatively automatic process whose outcome is not under one's control.

In the rest of this book I explore the hypothesis that there is a difference between theoretical and practical reasoning and that principles of revision can be distinguished from principles of reflection and from other principles of change in view. For the time being I simply assume that there is something about some changes in view that makes it reasonable to call these changes "instances of reasoning" and to call the relevant principles "rules of revision," distinguishing these changes from others that are significantly different. This assumption seems initially plausible. I will try to show that it is also fruitful.

Reasoning Distinguished from Argument or Proof

Reasoned change in view like Mary's does not seem to have been studied much except for some recent research into planning and "belief revision" in the field of artificial intelligence (e.g., Doyle 1980). One possible cause of this otherwise general neglect is that reasoning in this sense may often be conflated with reasoning in another sense, namely argument for, or proof of, a conclusion from premises via a series of intermediate steps.



Clearly, argument or proof is not at all the same sort of thing as reasoning in the sense of reasoned change in view. There is a clear difference in category. Rules of argument are principles of implication, saying that propositions (or statements) of such and such a sort imply propositions (or statements) of such and such other sort. Consider the following principle:

Modus Ponens: *P* and *if P then Q* taken together imply *Q*.

Such a rule by itself says nothing at all in particular about belief revision. It may be that some principles of belief revisions *refer* to such principles of argument, that is, to principles of implication. It is an important issue in the theory of reasoning, conceived as change in view, just how implication or argument may be relevant to reasoning. I discuss this issue at some length in chapter 2. My present point is simply to note that rules of argument are not by themselves rules for revising one's view.

This difference in category between rules of implication and, as we might say, rules of inference (rules of revision) lies behind other dif-

ferences between proof or argument and reasoning. For example, implication is *cumulative* in a way that inference may not be. In argument one accumulates conclusions; things are always added, never subtracted. Reasoned revision, however, can subtract from one's view as well as add to it. In order to express this point, the artificial intelligence work I have mentioned contrasts "monotonic reasoning," as in the usual sort of argument or proof, which is cumulative, with "nonmonotonic reasoning," as in ordinary reasoning or reasoned revision, which is not cumulative (Doyle 1980, 1982). But, although this terminology emphasizes the noncumulative character of reasoned revision, it is also potentially misleading in calling the ordinary sort of proof or argument "monotonic reasoning," because proof or argument is not of the same category as reasoned revision.

Induction and Deduction

Making a clear distinction between reasoning in the sense of reasoned change in view and reasoning in the sense of proof or argument can have a profound effect on how we view a variety of issues. For example, we might be led to question whether there are such things as "inductive arguments." These would be like "deductive arguments" except that the conclusion of an inductive argument would not have to follow logically from the premises, as in a deductive argument, but would only have to follow probabilistically (Black 1958).

Such inductive arguments would be "defeasible," that is, adding "premises" to an inductive argument might undercut the "validity" of the argument in a way that cannot happen with deductive arguments. For example, suppose that there is a valid or warranted inductive argument of the following form.

> Most *F*'s are *G*'s.
> *Y* is an *F*.
> So, *Y* is a *G*.

The "conclusion" here is not a deductive consequence of the premises; it can only be "made probable" by them. Adding premises in this case *can* undercut the argument. Suppose that we are given the following two additional premises:

> Most *FH*'s are not *G*'s.
> *Y* is an *H*.

Now the argument is the following:

> Most *F*'s are *G*'s.
> *Y* is an *F*.
> Most *FH*'s are not *G*'s.
> *Y* is an *H*.
> So, *Y* is a *G*.

At this point, it may well be that the conclusion is no longer made probable by all the premises; so the argument from the extended set of premises would not be inductively valid (Hempel 1960).

Rules of inductive argument would be rules of "inductive logic" as opposed to deductive logic. It happens, however, that there is no well-developed enterprise of inductive logic in the way that there is for deductive logic.

Now, why should we think there are inductive arguments and an inductive logic? It is clear enough that there is something that might be called inductive reasoning, that is, inductively reasoned change in view. But if we clearly distinguish reasoned change in view from argument, we cannot suppose that the existence of inductive reasoning by itself shows there is such a thing as inductive argument, nor can we suppose that it shows there is an inductive logic.

Indeed, if we clearly distinguish reasoning from argument, we cannot suppose that the existence of deductive arguments shows there is such a thing as deductive reasoning, that is, deductively reasoned change in view. As I have already observed, rules of deduction are rules of deductive argument; they are not rules of inference or reasoning. They are not rules saying how to change one's view. Nor (to anticipate the discussion of this issue in chapter 2) are they easily matched to such rules. Consider again modus ponens. This principle does not say that, if one believes *P* and also believes *if P then Q*, then one can infer *Q*, because that is not always so. Sometimes one should give up *P* or *if P then Q* instead.

Even if some sort of principle of belief revision corresponded to this logical principle, the principle of belief revision would have to be a different principle. For one thing, the logical principle holds without exception, whereas there would be exceptions to the corresponding principle of belief revision. Mary believes that if she looks in the cupboard, she will see a box of Cheerios. She comes to believe that she is looking in the cupboard and that she does not see a box of Cheerios. At this point, Mary's beliefs are jointly inconsistent and therefore *imply* any proposition whatsoever. This does not authorize Mary to *infer* any proposition whatsoever. Nor does Mary infer whatever she might wish to infer. Instead she abandons her first belief, concluding that it is false after all.

Furthermore, even before Mary fails to find any Cheerios in the cupboard, it would be silly for her to clutter her mind with vast numbers

of useless logical implications of her beliefs, such as *either she will have Cheerios for breakfast or the moon is made of green cheese.*

If there is a connection between standard principles of logic and principles of reasoning, it is not immediately obvious. There is a gap. We can't just state principles of logic and suppose that we have said something precise about reasoning. (I discuss the relation between logic and reasoning in chapter 2.)

Clearly, distinguishing between reasoning and argument can make one skeptical of the familiar idea that deduction and induction are different species of the same sort of thing. Obviously, there is deductive argument, but it is not similarly obvious that there is deductive reasoning. Again, it is not clear that there is such a thing as inductive argument, although we might say there is inductive reasoning. (It might be safer, however, to speak of theoretical reasoning instead of inductive reasoning, because theoretical reasoning contrasts with practical reasoning, which clearly exists, whereas to speak of inductive reasoning may suggest a contrast with deductive reasoning, which does not obviously exist.)

Analogous remarks also apply to the suggestion that there is such a thing as the practical syllogism (Anscombe 1957, p. 57). A syllogism is a form of argument, and although there is practical reasoning, there is not obviously any such thing as practical argument or logic and so not obviously any such thing as a practical syllogism.

Again, consider a defense of a "logic of entailment," which observes (1) in standard logic a contradiction logically implies any proposition at all, and (2) one is not justified in responding to the discovery that one's view is inconsistent by inferring anything whatsoever, concluding that (3) a new logic is needed (Meyer 1971). This line of thought loses plausibility if rules of inference or reasoning are distinguished from rules of implication or argument.

Finally, distinguishing reasoning from argument can make one worry that the work in artificial intelligence I have previously mentioned may be hampered by the so far unsuccessful search for principles of a nonmonotonic logic, in contrast to the usual principles of monotonic logic (McCarthy 1980; McDermott and Doyle 1980; Reiter 1980). It may be a mistake to expect principles of reasoning to take the form of a logic.

In short, distinguishing reasoning from argument can make one suspicious of certain arguments for inductive logic, practical syllogisms, a logic of entailment, and so on. It is unclear how work on such "logics" might contribute to the study of reasoned revision.

Descriptive versus Normative Theories

My aim in this book is to contribute to the development of a theory of reasoned revision, but I find it hard to say whether the theory I want is a *normative* theory or a *descriptive* theory. A normative theory says how people *ought* to reason, whereas a descriptive theory says how they actually *do* reason. The theory I envision tries to say either or both of these things.

Actually, normative and descriptive theories of reasoning are intimately related. For one thing, as we will see, it is hard to come up with convincing normative principles except by considering how people actually do reason, which is the province of a descriptive theory. On the other hand it seems that any descriptive theory must involve a certain amount of idealization, and idealization is always normative to some extent.

The distinction between a normative and a descriptive theory seems as clear as the thought that one might sometimes reason in a way in which one ought not to have reasoned, in which case there is something wrong with one's reasoning. So let us consider ways in which one can make mistakes while reasoning. There are at least four such ways:

1. One might start with false beliefs and by reasoning be led into further errors.
2. One might reach a conclusion that is perfectly "reasonable," even though it happens to be mistaken.
3. One can be careless or inattentive; one can forget about a relevant consideration or fail to give it sufficient weight; one can make mistakes in long division; one can fail to see something, to remember something, to attend carefully; and so on.
4. One can revise one's view in accordance with an incorrect rule of revision, thereby violating the correct rules.

Mistakes of type 1 or 2 do not seem to be errors of reasoning at all. Only mistakes of type 3 and 4 seem to be errors of reasoning. Mistakes of type 3 seem to be mistakes of "reflection," involving the violation of a maxim of reflection. Mistakes of type 4 would be mistakes of revision, involving the violation of a principle of revision.

I envision a theory that says something about principles of revision. One way to try to discover the *right* principles of revision might be to consider actual cases in which people make mistakes of type 4 to see why they are mistakes. Through seeing when the wrong principles are followed, we might hope to discover what the right principles are. But it is not easy to find cases in which people clearly change their views in accordance with incorrect principles of revision. It is difficult to come

up with an example that cannot be attributed instead to a mistaken belief, perhaps due to carelessness, so that the mistake is of type 1 and possibly type 3, rather than of type 4.

We cannot simply say one makes a mistake of type 4 whenever one reasons "fallaciously" in accordance with an "invalid" rule for changing one's view. If we distinguish clearly between argument and reasoning, we must agree that only arguments and proofs can be valid or invalid and that the notions of validity and invalidity have no clear application to changes in view, except in the sense that one can make a mistake about what validly implies what, a mistake that affects one's reasoning.

It is often said that there is a fallacy of "affirming the consequent," in which one reasons from the premises *if P then Q* and *Q* to the conclusion *P*. This would contrast with "affirming the antecedent," that is, modus ponens. But how are we to understand the contrast? Given a sharp distinction between reasoning and argument, we cannot suppose one's reasoning is valid if it proceeds in accordance with modus ponens and invalid if it proceeds in accordance with the principle of affirming the consequent. Modus ponens is a principle of argument or implication, not a principle of reasoned revision. If there is a "fallacy" here, it seems to involve making a type 1 mistake about what implies what.

Similar remarks apply to the so-called Gambler's Fallacy. This occurs, for example, in the game of roulette, in which one bets on where a spinning wheel with a pointer will stop. The pointer might end up on red or black (or very occasionally on green) and is equally likely to stop with the pointer on red as it is to stop with the pointer on black; that is, each time, the probability of red is the same as the probability of black. The Gambler's Fallacy consists in thinking that, under these conditions, red and black should each occur about half the time in any sufficiently long series of spins; so, if black has come up ten times in a row, red must be highly probable next time. This is a fallacy since it overlooks how the impact of an initial run of one color can become more and more insignificant as the sequence gets longer. If red and black occur each about half the time in a *long enough* sequence, they also occur about half the time in the somewhat longer sequence that includes ten extra occurrences of black at the beginning. For example, suppose red and black each occur 50% of the time in a sequence of 1000 spins of the wheel. Then in the longer sequence obtained by adding ten occurrences of black at the beginning, red and black each occur within half a percent of 50% of the time.

It may well turn out that all "fallacies" are best thought of either as mistakes of type 1, namely, reasoning from false beliefs in which the beliefs happen to be beliefs about what implies what or beliefs about probability, or as mistakes of type 3, involving carelessness or a failure to consider all the relevant possibilities. None of the fallacies clearly involves a distinctive mistake of type 4, in which a mistaken principle of change in view is followed. So, it seems that we cannot immediately use the existence of such fallacies to help us discover what the correct principles of revision are.

How then are we to begin to figure out what these principles of revision are? There seem to be two possible approaches. We can begin by considering how people actually *do* reason, by trying to figure out what principles they *actually follow*. Or we can begin with our "intuitions" as *critics* of reasoning. In either case we can then hope to find general principles. This will almost certainly involve some idealization. The suggested general principles will not coincide perfectly with our actual practice or with our intuitions about cases. This may lead us to modify the general principles, but it may also lead us to change our reasoning practice and/or our intuitions about what reasoning is correct. This can lead to a process of mutual adjustment of principles to practice and/or intuitions, a process of adjustment which can continue until we have reached what Rawls (1971) calls a reflective equilibrium. Furthermore, and this is important, we can also consider what rationale there might be for various principles we come up with and that can lead to further changes in principles, practices, and/or intuitions.

To repeat, even if we start by considering how people actually do reason, our account will probably have to involve a certain amount of idealization. Now, some kinds of idealization yield a normative theory, a notion of how one would reason if everything went right. So even this approach may yield a natural distinction between *is* and *ought*, between how things *do* happen and how they *ought* to happen. Indeed, it may do so in more than one way.

In what follows I consider matters from both the viewpoint of our intuitions as critics and the viewpoint of our actual practice. These two approaches yield somewhat different results. As we will see, the appeal to intuition tends toward a greater degree of idealization. In particular, it tends to overlook or minimize practical limitations, such as limitations on memory or on calculative capacity. What seems wrong when these limitations are not taken into account may be quite reasonable when they are taken into account. So the two approaches can seem at least initially to yield different results.

Human versus Artificial Reasoning

I am concerned with human reasoning, given the constraints of human psychology. Although occasionally I allude to work in artificial intel-

ligence, I am here concerned with this work not for its own sake but only for the light it may shed on human reasoning.

My conclusions about human reasoning may sometimes be relevant to work in artificial intelligence. For example, I argue in chapter 3 that people cannot do much probabilistic reasoning because of a combinatorial explosion such reasoning involves. If this is correct, the same limitation will apply to the "reasoning" of artificial intelligence systems. But human reasoning is affected by other limits to which artificial intelligence may not be subject, for example, limits on short-term memory. Seeing how humans reason in consquence of *these* limits may or may not be of much interest for artificial intelligence.

Summary

I am concerned with reasoned change in view. Such reasoning may involve giving up things previously accepted as well as coming to accept new things. I assume there is a difference between theoretical reasoning, which immediately modifies beliefs, and practical reasoning, which immediately modifies plans and intentions. I also assume we can distinguish maxims of reflection, saying what to think about before revising one's view, from principles of revision, the rules concerning the actual revision to be made.

Reasoning in the sense of reasoned change in view should never be identified with proof or argument; inference is not implication. Logic is the theory of implication, not directly the theory of reasoning. Although we can say there is inductive reasoning, it is by no means obvious that there is any such thing as inductive argument or inductive logic. Nor does the existence of practical reasoning show there is such a thing as a practical syllogism or a practical logic.

Finally, it is not at this point easy to distinguish a descriptive theory of reasoned revision from a normative theory. Any normative investigation must begin by considering how people actually do reason and how people criticize reasoning. Any descriptive theory has to make use of idealization.

Chapter 2

Logic and Reasoning

Even if they agree that logic is not by itself a theory of reasoning, many people will be inclined to suppose that logic has some sort of special relevance to the theory of reasoning. In this chapter I argue that this inclination should be resisted. It turns out that logic is not of any special relevance.

Implications, Inconsistency, and Practical Limits

If logic does have special relevance to reasoning, it would seem that its relevance must be captured at least roughly by the following two principles.

> *Logical Implication Principle* The fact that one's view logically implies P can be a reason to accept P.
>
> *Logical Inconsistency Principle* Logical inconsistency is to be avoided.

These are distinct principles. Suppose one believes both P and also *if* P *then* Q. Since these beliefs imply Q, the Logical Implication Principle says this may give one a reason to believe Q. It does not say one should also refrain from believing Q's denial, *not* Q. Believing *not* Q when one also believes P and *if* P *then* Q is contrary to the Logical Inconsistency Principle, not to the Logical Implication Principle. On the other hand the Logical Inconsistency Principle does not say one has a reason to believe Q given that one believes P and *if* P *then* Q.

Neither principle is exceptionless as it stands. Each holds, as it were, other things being equal. Each is defeasible. For example, the Logical Implication Principle entails that, if one believes both P and *if* P *then* Q, that can be a reason to believe Q. But, clearly, that is not *always* a reason to believe Q, since sometimes when one believes P and also believes *if* P *then* Q, one should *not* come to believe Q. Remember Mary who came to believe three inconsistent things: If she looks in the closet she will see a box of Cheerios, she is looking in the closet,

but she does not see a box of Cheerios. Mary should not at this point infer that she does see a box of Cheerios from her first two beliefs.

This suggests modifying the Logical Implication Principle:

> *Logical Closure Principle* One's beliefs should be "closed under logical implication." In other words there is something wrong with one's beliefs if there is a proposition logically implied by them which one does not already believe. In that case one should either add the implied proposition to one's beliefs or give up one of the implying beliefs.

But the Logical Closure Principle is not right either. Many trivial things are implied by one's view which it would be worse than pointless to add to what one believes. For example, if one believes P, one's view trivially implies "either P or Q," "either P or P," "P and either P or R," and so on. There is no point in cluttering one's mind with all these propositions. And, of course, there are many other similar examples.

Here I am assuming the following principle:

> *Clutter Avoidance* One should not clutter one's mind with trivialities.

This raises an interesting issue. To suppose one's mind could become cluttered with beliefs is to suppose such things as (1) that it takes time to add to one's beliefs further propositions that are trivially implied by them, time that might be better spent on other things, and/or (2) that one has "limited storage capacity" for beliefs, so there is a limit on the number of things one can believe, and/or (3) that there are limits on "information retrieval," so the more one believes the more difficult it is to recall relevant beliefs when one needs them.

Such suppositions presuppose that beliefs are explicitly "represented" in the mind in the sense that these representations play the important role in perception, thought, and reasoning that we think beliefs play.

But we must be careful in stating this presupposition. Not all one's beliefs can be explicitly represented in this way, since then one could believe only finitely many things. But one can and does believe infinitely many things. For example, one believes the earth does not have two suns, the earth does not have three suns, the earth does not have four suns, and so on.

In order to accommodate this point, I assume that we can distinguish what one believes *explicitly* from what one believes only *implicitly*. Then we can take the principle of clutter avoidance to apply to what one believes explicitly.

Explicit and Implicit Belief

I assume one believes something explicitly if one's belief in that thing involves an explicit mental representation whose content is the content of that belief. On the other hand something is believed only implicitly if it is not explicitly believed but, for example, is easily inferable from one's explicit beliefs. Given that one explicitly believes the earth has exactly one sun, one can easily infer that the earth does not have two suns, that the earth does not have three suns, and so on. So all these propositions are things one believes implicitly.

That is an example in which implicit beliefs are implied by explicit beliefs. There are also cases in which one implicitly believes something that is easily inferable from one's beliefs without being strictly implied by them. An example might be one's implicit belief that elephants don't wear pajamas in the wild (Dennett 1978).

There is also another way in which something can be implicitly believed—it may be implicit in one's believing something else. For example, in explicitly believing P, it may be that one implicitly believes one is justified in believing P. The proposition that one is justified in believing P is not ordinarily implied by the proposition P and may not be inferable from one's explicit beliefs, but it may be that in believing P one is committed to and so implicitly believes the proposition that one is justified by believing P. (I discuss this and related possibilities in chapter 5.)

It is a possible view that *none* of one's beliefs are explicit, that is, that none are explicitly represented and all are only implicit in one's mental makeup. This is a form of behaviorism about belief. There is surprisingly much that can be said in favor of such behaviorism (Dennett 1978; Stalnaker 1984), but I suppose that whatever is ultimately the right view of belief must allow that unbridled inference can lead to too much clutter either in what one explicitly *believes* or in whatever explicit thing underlies belief. Therefore I ignore the possibility of such behaviorism and continue to assume that one's implicit beliefs are implicit in one's believing certain things explicitly. If this is wrong, I doubt that it is so wrong as to affect the conclusions I draw from this assumption.

In this connection it might be useful for me to digress briefly to observe that the distinction between explicit and implicit belief is not the same as either the distinction between belief that is available to consciousness and unconscious belief or that between "occurrent" and "dispositional" beliefs.

We normally consider a belief "unconscious" if one is not aware one has it and one cannot easily become aware of it simply by considering

whether one has it. Otherwise the belief is available to consciousness. Now, clearly, implicit beliefs can be available to consciousness. The belief that the earth does not have two suns is normally only implicit in one's explicit beliefs and is not itself explicitly represented, even though it is immediately available to consciousness in the sense that, if one considers whether one believes it, one can immediately tell one does.

On the other hand a belief can be explicitly represented in one's mind, written down in Mentalese as it were, without necessarily being available to consciousness. For example, one might explicitly believe that one's mother does not love one, even though this belief may not be consciously retrievable without extensive psychoanalysis. So the distinction between implicit and explicit beliefs is not the same as that between unconscious beliefs and those available to consciousness.

Turning now to the distinction between occurrent and dispositional beliefs, we can say a belief is occurrent if it is either currently before one's consciousness or in some other way currently operative in guiding what one is thinking or doing. A belief is merely dispositional if it is only potentially occurrent in this sense. Any merely implicit belief is merely dispositional, but explicit beliefs are not always occurrent, since only some explicit beliefs are currently operative at any given time. So the distinction between implicit and explicit beliefs is not the same as that between occurrent and dispositional beliefs.

So much for this digression comparing these various distinctions in kinds of beliefs.

Let me return to the discussion of the Logical Closure Principle, which says one's beliefs should be closed under logical implication. Clearly this principle does not apply to explicit beliefs, since one has only a finite number of explicit beliefs and they have infinitely many logical consequences. Nor can the Logical Closure Principle be satisfied even by one's implicit beliefs. One cannot be expected even implicitly to believe a logical consequence of one's beliefs if a complex proof would be needed to see the implication.

It won't help to change the Logical Closure Principle to say one's beliefs should be closed under *obvious* logical implication. That would come to the same thing, since any logical implication can eventually be demonstrated by a proof consisting entirely of a series of obvious steps. This means that, if beliefs are required to be closed under obvious logical implication, they are required to be closed under any logical implication, obvious or not. So, since beliefs cannot be required to be closed under logical implication, they cannot be required to be closed under obvious logical implication either.

Clutter Avoidance

How is the principle of clutter avoidance to be used? It seems absurd for it to figure explicitly in one's reasoning, so that one refrains from drawing an otherwise acceptable conclusion on the grounds of clutter avoidance. Once one is explicitly considering whether or not to accept a conclusion, one cannot decide not to on such grounds. One might rationally decide not to try to remember it, perhaps, but one cannot decide not to believe it at least for the moment.

Suppose George is trying to convince Bob that *P*. George shows how *P* is a deductive consequence of things Bob believes. Bob accepts the validity of George's argument and refuses to change his belief in any of the premises, but he also refuses to accept the conclusion *P*, citing clutter avoidance as his reason for refusing. That is absurd. (I am indebted to Robert Stalnaker for this example.)

But the Principle of Clutter Avoidance is not just a principle about what one should try to remember. It would be a violation of clutter avoidance if one spent all one's time thinking up trivial consequences of one's beliefs even if one refrained from committing these consequences to memory. In that case one would be cluttering up one's short-term processing capacities with trivialities.

The Principle of Clutter Avoidance is a metaprinciple that constrains the actual principles of revision. The principles of revision must be such that they discourage a person from cluttering up either long-term memory or short-term processing capacities with trivialities. One way to do this would be to allow one to accept a new belief *P* only if one has (or ought to have) an interest in whether *P* is true. (This is discussed in chapter 6.)

Unavoidable Inconsistency and the Liar Paradox

I was saying that neither the Logical Implication Principle nor the Inconsistency Principle is without exception. I have indicated why this is so for the Logical Implication Principle, which says one has a reason to believe the logical implications of one's beliefs. Similar remarks hold for the Logical Inconsistency Principle, which says one should avoid inconsistency.

To see that the Logical Inconsistency Principle has its exceptions, observe that sometimes one discovers one's views are inconsistent and does not know how to revise them in order to avoid inconsistency without great cost. In that case the best response may be to keep the inconsistency and try to avoid inferences that exploit it. This happens in everyday life whenever one simply does not have time to figure out

what to do about a discovered inconsistency. It can also happen on more reflective occasions. For example, there is the sort of inconsistency that arises when one believes that not all one's beliefs could be true. One might well be justified in continuing to believe that and each of one's other beliefs as well.

There are also famous logical paradoxes. For example, the liar paradox involves reflection on the following remark, which I call (*L*):

(*L*) is not true.

Thinking about (*L*) leads one into contradiction. If (*L*) is not true, things are as (*L*) says, so (*L*) must be true. But if (*L*) is true, then it is true that (*L*) is not true, so (*L*) must not be true. It seems (*L*) is true if and only if (*L*) is not true. But that is a contradiction.

The paradox arises from our uncritical acceptance of the following:

Biconditional Truth Schema "*P*" is true if and only if *P*.

To see that this schema is indeed the culprit, notice that one instance of it is

"(*L*) is not true" is true if and only if (*L*) is not true.

Since (*L*) = "(*L*) is not true," this instance is equivalent to the self-contradictory

(*L*) is true if and only if (*L*) is not true

Various restrictions on the Biconditional Truth Schema have been suggested in order to avoid the liar paradox, but none is completely satisfactory (Kripke 1975, Herzberger 1982). So, the rational response for most of us may simply be to recognize our beliefs about truth are logically inconsistent, agree this is undesirable, and try not to exploit this inconsistency in our inferences. (The danger is that, since inconsistent beliefs logically imply anything, if one is not careful, one will be able to use this fact to infer anything whatsoever.)

In practice the best solution may be to retain the Biconditional Truth Schema and yet avoid contradiction by interpreting the Schema not as something that holds without exception but rather as something that holds "normally" or "other things being equal." It is then a "default assumption." One accepts any given instance of the Biconditional Truth Schema in the absence of a sufficiently strong reason not to accept it. One does not apply the Schema to (*L*) because doing so leads to contradiction.

This does not seem to be a satisfactory solution from the point of view of logic, since we take logic to require precise principles with precise boundaries, not principles that hold merely "normally" or "other things being equal." But in ordinary life we accept many principles of this vaguer sort.

My point about the Logical Inconsistency Principle remains. One may find oneself with inconsistent beliefs and not have the time or ability to trace the sources of the inconsistency (e.g., the Biconditional Truth Schema). In that event, it is rational simply to retain the contradictory beliefs, trying not to exploit the inconsistency.

Immediate Implication and Immediate Inconsistency

I turn now to an issue about the Logical Implication and Inconsistency Principles I have so far mentioned only in passing. One might have *no* reason to accept something that is logically implied by one's beliefs if there is no short and simple argument showing this. To take an extreme example, one accepts basic principles of arithmetic that logically imply some unknown proposition *P* which is the answer to an unsolved mathematical problem: but one has no reason to believe *P* if one is not aware that *P* is implied by these basic principles. This suggests revising the Logical Implication Principle:

Recognized Logical Implication Principle One has a reason to believe *P* if one *recognizes* that *P* is logically implied by one's view.

Similarly, we might revise the Logical Inconsistency Principle:

Recognized Logical Inconsistency Principle One has a reason to avoid believing things one recognizes to be logically inconsistent.

However, there is a problem with this. It would seem one can *recognize* a logical implication or logical inconsistency only if one has the relevant concept of logical implication or logical inconsistency. But it would seem that few people have such concepts, at least if this involves distinguishing logical implication and inconsistency from other sorts of implication and inconsistency. Consider the following examples:

P or *Q* and *not P* taken together imply *Q*.
$A = B$ and $B = C$ taken together imply $A = C$.
$A < B$ and $B < C$ taken together imply $A < C$.
A is part of B and *B is part of C* taken together imply *A is part of C*.
X is Y's brother implies *X is male*.
Today is Thursday implies *Tomorrow is Friday*.
X plays defensive tackle for the Philadelphia Eagles implies *X weighs more than 150 pounds*.

People who recognize these and related implications do not in any consistent way distinguish them into purely logical implications and

others that are not purely logical. (Only the first counts as purely logical in "classical" first-order predicate logic without identity. Sometimes principles for identity are included as part of logic, in which the second also counts as a logical implication.) So the Recognized Logical Implication and Inconsistency Principles would seem to have only a limited application.

To some extent this objection can be met by generalizing the principles, dropping specific mention of logical implication and inconsistency. Then the principles would be stated as follows:

> *Recognized Implication Principle* One has a reason to believe P if one recognizes that P is implied by one's view.

> *Recognized Inconsistency Principle* One has a reason to avoid believing things one recognizes to be inconsistent.

These principles still apply only to people who have concepts of implication and inconsistency. But this is not so clearly problematical, if only because it is not clear what it takes to have these concepts.

I suggest it is enough to be able to make reasoned changes in one's view in a way that is sensitive to implication and inconsistency. Someone who is disposed to treat beliefs in P and *if* P *then* Q as reasons to believe Q has, by virtue of that very disposition, an appropriate ability to recognize this sort of implication, at least if this disposition is also accompanied by the disposition not to believe P, *if* P *then* Q, and *not* Q. And the latter sort of disposition might reflect an appropriate ability to recognize that sort of inconsistency.

I am inclined to take as fundamental certain dispositions to treat propositions in certain ways, in particular, dispositions to treat some propositions as *immediately implying* others and some as *immediately inconsistent* with each other. That is, I am inclined to suppose that the basic notions are

$P, Q, \ldots, R$ immediately imply S for A

and

$P, Q, \ldots, R$ are immediately inconsistent for A.

It is unclear to me whether these notions can be reduced to others in any interesting way. (I am indebted to Scott Soames for raising this issue.) If A is disposed to treat some beliefs as implying others, then A is disposed to treat beliefs that immediately imply something as giving him or her a reason to believe that thing; but we cannot *identify*

$P, Q, \ldots, R$ immediately imply S for A

with

A is disposed to treat $P, Q, \ldots, R$ as a reason to believe S.

For one thing, A's general disposition may be overridden by other considerations in a particular case, for example, if S is absurd. In that case, some of A's beliefs will immediately imply a particular proposition for A, although A is not disposed to treat those beliefs as reasons for believing that proposition. Furthermore, beliefs can be treated by A as *reasons* for believing a conclusion even though A does not take those beliefs to *imply* that conclusion.

Similarly, we can say A is disposed to avoid believing things that are immediately inconsistent for A but we cannot identify a set of beliefs' being inconsistent for A with A's having a disposition to avoid believing all the members of that set. On the one hand the general disposition may be overridden in a particular case, as when A is disposed to believe the premises of the liar paradox. On the other hand (as Soames observes) there is Moore's paradox: One is strongly disposed not to believe both P and that one does not believe P while realizing that these propositions are perfectly consistent with each other.

So, I am inclined simply to assume one has certain basic dispositions to take some propositions immediately to imply other propositions and to take some propositions as immediately inconsistent with each other. More generally, I assume one can have general dispositions with respect to certain patterns of immediate implication and inconsistency even if some instances of the patterns are so long or complex or otherwise distracting that one has no particular disposition to take those particular instances to be immediate implications or inconsistencies.

Summary and Conclusion

I began by suggesting logic might be specially relevant to reasoning in two ways, via implication and inconsistency. It seemed the relevant principles would be defeasible, holding only other things being equal. Furthermore, they would apply only to someone who recognized the implication or inconsistency. Since this recognition might be manifested simply in the way a person reacts to these implications and inconsistencies, I suggested that certain implications and inconsistencies are "immediate" for a given person. (In appendix A, I discuss whether basic logical notions can be defined in terms of such immediate implications and inconsistencies. I also discuss the hypothesis that only logical implications and inconsistencies are immediate and that others are mediated by the acceptance of certain nonlogical principles, concluding that this hypothesis may be impossible to refute.)

My conclusion is that there is no clearly significant way in which *logic* is specially relevant to reasoning. On the other hand immediate *implication* and immediate *inconsistency* do seem important for reasoning, and so do implication and inconsistency. Sometimes, reasoning culminates in the conclusion that a certain argument is a good one or that certain propositions are inconsistent. But that is not to say that logical implication or logical inconsistency has any special status in human reasoning.

Chapter 3

Belief and Degree of Belief

Probabilistic Implication

We have a rule connecting implication and reasoning:

> *Principle of Immediate Implication* That P is immediately implied by things one believes can be a reason to believe P.

Is there also a weaker probabilistic version of this rule?

> *Hypothetical Principle of Immediate Probabilistic Implication* That P is obviously highly probable, given one's beliefs, can be a reason to believe P.

Suppose Mary purchases a ticket in the state lottery. Given her beliefs, it is obviously highly probable that her ticket will not be one of the winning tickets. Can she infer that her ticket will not win? Is she justified in believing her ticket is not one of the winning tickets?

Intuitions waver here. On the one hand, if Mary is justified in believing her ticket is not one of the winning tickets, how can she be justified in buying the ticket in the first place? Furthermore, it certainly seems wrong to say she can *know* that her ticket is not one of the winning tickets if it is really a fair lottery. On the other hand the probability that the ticket is not one of the winning tickets seems higher than the probability of other things we might easily say Mary knows. We ordinarily allow that Mary can come to know various things by reading about them in the newspaper, even though we are aware that newspapers sometimes get even important stories wrong.

This issue is one that I will return to several times, but I want to begin by considering a suggestion which I think is mistaken, namely, that the trouble here comes from not seeing that belief is a matter of degree.