2
THE ARCHITECTURE OF EPISTEMIC COGNITION

1. Reasoning and Q&I Modules

Epistemic cognition is concerned with updating and maintaining our beliefs about the world. Epistemology has traditionally focussed on epistemic *reasoning*, but it is important to realize that many beliefs are formed by processes that can be usefully contrasted with stereotypical reasoning. What we normally think of as reasoning is a step-by-step process whereby we form beliefs one after another, each based upon previous beliefs. As such, reasoning is an essentially serial process. Whenever possible, human cognition is accelerated by employing our inherently slow hardware in parallel processing. Unfortunately, much of reasoning cannot be done in parallel, so human cognition includes other processes that also issue in beliefs and actions. For instance, a human being does not have to pull out a pocket calculator and compute trajectories in order to catch a baseball or get out of the way of a bus. We have a built-in mechanism that allows us to estimate such trajectories very quickly. This seems equally desirable in artificial agents. Even though they may be based upon faster hardware, parallel processing is still going to be quicker than serial processing.

Non-ratiocinative processes achieve their speed by building in assumptions about the environment, and then hardwiring computations appropriate to those assumptions. For instance, we can predict the trajectories of flying objects very quickly, but only by assuming that nothing will interfere with their flight. If a baseball is going to bounce off a telephone pole, we must wait until we see its new direction of flight before we can predict where it is going to land. I will refer to these non-ratiocinative processes as *Q&I modules* ("quick and inflexible" modules). It must not be supposed that human Q&I modules are concerned exclusively with motor skills. Psychological evidence strongly suggests that most everyday inductive and probabilistic inference is carried out by Q&I modules.[1]

The advantage of Q&I modules is speed. The advantage of reasoning, on the other hand, is extreme flexibility. It seems that reasoning can in principle deal with any kind of situation, but it is slow. In complicated situations we may have no applicable Q&I modules, in which case we have no choice but to undertake explicit reasoning about the situation. In other cases, human beings accept the output of the Q&I modules *unless* they have some explicit reason for not doing so. Reasoning is used to monitor the output and override it when necessary. In sum, the role of reasoning should be (1) to deal with cases to which Q&I modules do not apply and (2) to monitor and override the output of Q&I modules as necessary. A rational agent can be viewed as a bundle of Q&I modules with reasoning sitting on top and tweaking the output as necessary.

To fall within the purview of rationality, a cognitive process must be introspectible. The course of reasoning is introspectible, but Q&I modules are not. Q&I modules are black-boxes. All we can introspect is the output, which is a belief. Accordingly, Q&I modules do not fall within the purview of rationality. The interaction between reasoning and Q&I modules is introspectible, however. A cognitive agent could be devoid of reasoning. Its cognition could consist entirely of Q&I modules. It is quite possible that many moderately sophisticated animals work in this way. However, if a cognitive agent is equipped with both reasoning and Q&I modules, then it seems clear that reasoning should trump the Q&I modules. As indicated above, one of the main points of reasoning is to ride herd on the Q&I modules, correcting their output

---

[1] Tversky and Kahneman [1974].

as necessary. So one characteristic of rationality is that reasoning should have total overall control. Because the interaction between reasoning and Q&I modules is introspectible, there can be rational norms to this effect.

Giving Q&I modules a subservient role is not to denigrate their importance. It is doubtful that any agent could survive in a hostile world relying upon reasoning as its exclusive means of cognition. But in rational agents, reasoning must provide the court of last appeal. Given time to make the appeal, the agent will direct its activity in accordance with the pronouncements of reasoning insofar as these differ from the pronouncements of its Q&I modules.

2. Epistemic Reasoning and Epistemic Norms

Reasoning is a form of rational cognition, and as such it is governed by rational norms. The norms governing specifically epistemic reasoning are epistemic norms, and their description is one of the main interests of epistemology. One of the main purposes of this book is to give a precise account of some of the more important epistemic norms involved in human rationality. The primary focus will be human rationality, but the results will be of more general significance. This is for two reasons. First, it is hard to build a cognitive system that accomplishes the kind of sophisticated epistemic cognition humans do. The easiest way to go about it is to use the human model as the guiding inspiration for the construction of artificial agents. Second, it can (and will) be argued that many aspects of human epistemic cognition represent the *only* way to solve various logical and computational problems involved in cognitive engineering, and as such other agents *must* replicate human epistemic norms in those respects. It is noteworthy that a number of purely logical arguments can be found in the epistemology literature to the effect that human rational thought could not work in certain ways—that would be logically impossible. Insofar as those arguments are correct, they show equally that other rational agents could not work that way either. So they bear on more than human epistemology. They bear upon the epistemology of arbitrary rational agents.

The epistemological literature contains numerous proposals for the contents of epistemic norms. In evaluating these proposals, we must keep in mind the distinction between procedural epistemology and descriptive epistemology, and more specifically the distinction between epistemological theories aimed at describing the norms governing procedural epistemic justification and the epistemological theories aimed at solving the Gettier problem. Only the former are directly relevant to the current enterprise.

2.1 Internalist and Externalist Theories

There is a distinction in epistemology between *internalist* and *externalist* theories. Internalist theories are those according to which epistemic norms must be internalist, and internalist norms are those appealing only to internal states of the cognizer in determining how cognition should proceed. Externalist theories endorse epistemic norms appealing to external considerations. The best known externalist theories are versions of reliabilism. Reliabilist epistemic norms appeal to the reliability of cognitive processes, recommending those that are *in fact reliable* (not those that are believed by the cognizer to be reliable) and censuring others.

It takes little reflection to realize the externalist theories cannot be correct theories of procedural epistemic justification, because they are not the kind of norms that could direct cognition. We might have a norm appeal to the agent's *beliefs* about whether a cognitive process is reliable, but that would be an internalist norm. It appeal to beliefs, which are internal states. By contrast, a reliabilist norm appeals to whether cognitive processes are in fact reliable, regardless

of what the agent beliefs about their reliability. There is no way that a system of cognition can be responsive to de facto reliability. Consequently, such a norm is a nonstarter in a procedural theory.

The unmistakable conclusion is that reliabilism should not be viewed as a procedural epistemological theory. It is about something else—presumably the analysis of knowledge. Correct epistemic norms have to be internalist. And note that this is true for any cognitive agents, not just human beings. Let us turn then to an examination of internalist epistemological theories.

Epistemological theories can be classified in several ways.[2] We have discussed the internalism/externalism distinction, and rejected all externalist theories. Within the class of internalist theories, we can distinguish between *doxastic* and *nondoxastic* theories. Doxastic theories are those endorsing the "doxastic assumption", according to which epistemic norms can appeal only to what beliefs the agent has. Thus, for example, a doxastic theory might (somewhat simplistically) endorse a norm to the effect that if the agent believes P and believes $(P \rightarrow Q)$ then the agent should come to believe Q. Stereotypical examples of reasoning involve reasoning from beliefs to beliefs, so doxastic theories model all epistemic norms on norms governing such reasoning.

Epistemologists have often made the doxastic assumption almost without noticing it. When they did notice it, they justified it with a simple rationale: all our information about the world is encapsulated in beliefs. It seems that in deciding what to believe, we *cannot* take account of anything except insofar as we have beliefs about it. Consequently, nothing can enter into our epistemic norms except our beliefs.

Doxastic theories can in turn be divided into two groups—foundations theories and coherence theories.

2.2 Foundations Theories

Foundations theories are distinguished by the view that knowledge has "foundations". The foundations theorist begins with the psychological observation that all knowledge comes to us through our senses. Our senses provide our only contact with the world around us. Our simplest beliefs about the world are in direct response to sensory input, and then we reason from those simple beliefs to more complicated beliefs (for example, inductive generalizations) that cannot be acquired on the basis of single instances of sense perception. This psychological picture of belief formation suggests a parallel philosophical account of epistemic justification according to which those simple beliefs resulting directly from sense perception form an epistemological foundation and all other beliefs must be justified ultimately by appeal to these *epistemologically basic beliefs*. The basic beliefs themselves are not supposed to stand in need of justification. They are in some sense "self-justifying". One is automatically justified in such a belief merely by virtue of having it. Making this precise, a *foundations theory* is any doxastic theory proposing that some privileged subclass of beliefs has a special status in determining what other beliefs are justified.

The main problem for foundations theories lies in finding suitable beliefs to serve as the epistemologically basic foundation. Basic beliefs must be *self-justifying* in the sense that one can be justified in holding such a belief merely by virtue of the fact that one does hold it—one does not need an independent reason for holding a basic belief. This is captured by the concept of a *prima facie justified belief*:

> A belief is *prima facie justified* for a person S if and only if it is only possible for S to hold the belief unjustifiedly if he has reason for thinking he should not hold the belief (equivalently, it

[2] I follow the taxonomy of Pollock [1987].

is necessarily true that if S holds the belief and has no reason for thinking he should not then he is justified in holding the belief).

Foundational beliefs must be at least prima facie justified. Foundationalists have generally thought they had an even stronger status, being either *incorrigly justified* or *incorrigible*:

A belief is *incorrigibly justified* for a person S if and only if it is impossible for S to hold the belief but be unjustified in doing so.

A proposition P is *incorrigible* for a person S if and only if (1) it is necessarily true that if S believes P then P is true, and (2) it is necessarily true that if S believes ~P then P is false.

(Incorrigibility is defined this way to get around the problem of necessary beliefs.)

It is clear that most beliefs are neither incorrigible nor self-justifying. For example, if I believe I see an airplane in the sky, I might be wrong—it might be an eagle. So this belief is not incorrigible. Furthermore, I could believe it for a bad reason, e.g., a hasty generalization, in which case it would not be justified. Thus merely holding the belief does not make it justified, not even *prima facie justified*.

Traditional epistemologists were quick to recognize that most ordinary beliefs not suited for being epistemologically basic. Instead, they seized upon "appearance beliefs" as their candidate for epistemologically basic beliefs. Vision researchers generally agree that our perceptual apparatus is computationally complex, producing a sequence of (nonintrospectible) mental states representing successive stages of the computation. It begins with a two-dimensional array of stimulation on the retina, which is then parsed into shapes, textures, spatial relations, and so on, and finally produces beliefs about the world. We can diagram it crudely as in figure 2.1. At some point in the processing, a belief is produced. The immediately preceding state is presumed to be some kind of visual image. We will call it a *percept*. The proposal of the traditional epistemologist was then that the basic belief is a belief to the effect that the agent has the percept. In othe words, "There is a red object before me" cannot be epistemologically basic, but the proposal is that "I am having a visual experience of a putative red object" is. These are what are called "appearance beliefs".

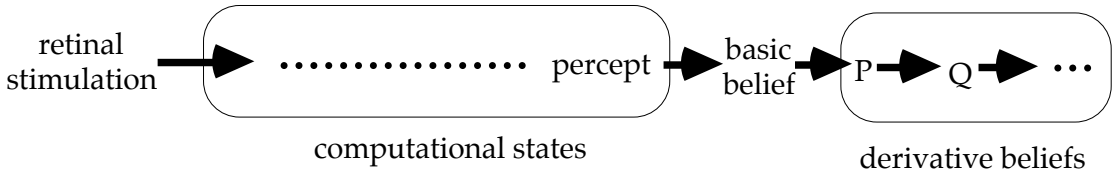

Figure 2.1 Perceptual processing

2.2.1 Are appearance beliefs epistemologically basic?

The traditional foundationalist proposes that (1) appearance beliefs are epistemologically basic, and (2) all other justified beliefs are supported by reasoning from appearance beliefs. Unfortunately, both claims are false. First, let me propose a counterexample to the claim that appearance beliefs are epistemologically basic. Because shadows on white surfaces are normally grey, most people think that shadows on snow are grey. But a discovery made fairly early by every landscape painter is that they are actually blue. A person having the general belief that shadows on snow are grey may, when queried about how a particular snow-shadow looks to him, reply that it looks grey, without paying any serious attention to his percept. His belief

about how it looks is based upon his general belief rather than inspection of his percept, and is accordingly wrong. This shows that the belief is not incorrigible. Furthermore, should his general belief itself be unjustifed, then his belief about how the shadow looks will also be unjustified, establishing that it is neither incorrigibly justified nor prima facie justified. This seems to show that beliefs about how things look to us do not have the logical properties minimally required to serve as foundational beliefs.

There is a response to this counterexample that has considerable intuitive pull, at least initially. This is to agree that not all beliefs about how things appear to us are prima facie justified, but those *based upon actually being appeared to in that way* are. Taken literally, this makes no sense. Prima facie justification is a logical property of propositions. A proposition cannot have such a property at one time and fail to have it at another. But the claim actually being made here is presumably a different one, viz., that when we are appeared to in a certain way, that in and of itself can make us at least defeasibly justified in believing that we are appeared to in that way. I think that this is true. But notice that a theory providing such a foundation for epistemic justification is no longer a doxastic theory. The justifiedness of beliefs is no longer determined exclusively by *what* we believe. What percepts we have is also relevant. Thus a theory proposing this is not a foundations theory in the sense defined above.

2.2.2 Are there enough appearance beliefs?

A second general argument against foundationalism has a simple structure. If all of our justified beliefs are inferred in some way from beliefs about percepts, then we must have *enough* beliefs about percepts to make that possible. Unfortunately, it does not seem that we do. People rarely have any beliefs at all about how things look to them. When I look around the room, I form all sorts of beliefs about physical objects, people, etc., but I rarely have any thoughts at all about how things look to me. Of course, I can quite easily acquire such beliefs by turning my attention inwards, but that requires a deliberate change of attention and is not something that we ordinarily do when perceiving the world around us. In normal cases of perceptual knowledge, it seems we move directly from our percepts to our beliefs about the world, without going through intermediary beliefs about how things appear to us.

Epistemologists and philosophers of mind might try to escape this objection by appealing to the distinction between occurrent and non-occurrent beliefs, and insisting that although we do not have occurrent beliefs about our percepts, we do have non-occurrent beliefs, and they provide the epistemological foundation required by foundationalist theories. Philosophers often talk about non-occurrent beliefs, but the concept is far from being a clear one. Sometimes non-occurrent belief is identified with dispositional belief. That is, we have a non-occurrent belief that P iff we have a disposition to form the occurrent belief that P when we consider the matter. Other philosophers have proposed that we non-occurrently believe all the infinitely many logical consequences of our occurrent beliefs. Whatever we are to make of non-occurrent beliefs, they are by definition not presently occurring mental states. But that seems to imply that they cannot play an active role in cognition, and cannot be among the internal states to which rational norms can appeal. So this is not a way of salvaging foundationalism.

2.2.3 Other Rational Architectures

The reasons given above for rejecting foundationalism only supports its rejection as a description of human rational cognition. Humans do not work that way, but there is no apparent reason why there could not be cognitive agents with a somewhat different cognitive architecture that conforms to traditional foundationalism. Presumably, the reason humans are constructed so that they move directly from percepts to beliefs about the physical world is that it is usually unnecessary to form beliefs about percepts. We can form them when they are useful, but they are not usually useful, and having to form them in all cases would expend our limited cognitive resources unnecessarily.

2.3 Coherence Theories

Coherence theories are defined to be any doxastic theories that are not foundations theories. In other words, coherence theories allege that epistemic norms can appeal only to beliefs, but they do not give any special subset of beliefs a privileged status. There are no epistemologically basic beliefs.

2.3.1 Positive and Negative Coherence Theories

We can distinguish between two kinds of coherence theories. *Negative coherence theories* take all beliefs to be prima facie justified. That is, you do not require reasons in order to be justified in holding a belief. For any belief, you are automatically justified in holding it unless you have reasons for thinking you should not. This is analogous to a foundations theory in which all beliefs are basic. *Positive coherence theories* go the other way and rule that *no* beliefs are basic—for any justified belief, you need reason for believing. Positive coherence theories assign no positive role to reasoning. Reasoning is not relevant to *producing* beliefs, only to rejecting them. All reasoning can do is reveal reasons for *not* holding beliefs.

As a theory of human rationality, negative coherence theories are clearly wrong. Beliefs are not automatically justified just by virtue of having them. Suppose a belief just pops into Jones' head to the effect that there are 312,642 redheads living in New Zealand. Jones has no reason to think this is true, but he doesn't really know much about New Zealand, so he has no reason to think it false either. We would not regard him as justified in holding this belief, despite the fact that he does not have any other belief that constitutes a reason for not holding it.

It is harder to argue that you couldn't have a rational agent whose rational architecture was properly described by a negative coherence theory. Humans do sometimes have beliefs just pop into their heads, often on the basis of wishful thinking, but an agent with a more controlled system of belief generation might never acquire such spurious beliefs. Then it is not clear that there would be anything wrong with regarding all of its beliefs as prima facie justified. It is hard to see how such an architecture would be inferior to the human rational architecture. This is related to another very interesting question. Why are humans so constructed that they *can* hold unjustified beliefs? For instance, we engage in wishful thinking, but if we were designing a cognitive agent from scratch, wouldn't it be better to build it so that it couldn't engage in wishful thinking? Why not just build the agent so that its epistemic norms are laws rigidly governing its cognition rather than just being procedural norms that can be violated? There may be a good answer to this question, but at this point I do not know what it is. So at this point, all that can be said with confidence is that the human rational architecture is not properly described by a negative coherence theory.

2.3.2 Linear and Holistic Coherence Theories

Within positive coherence theories, we can subdivide the theories further in terms of the role they assign to reasoning. *Linear positive coherence theories* characterize epistemic justification in terms of a *reason-for* relation analogous to that employed in foundations theories. It is a relation between a conclusion and a single belief or small set of beliefs providing a reason for the conclusion. By contrast, *holistic positive coherence theories* characterize epistemic justification directly in terms of structural features of the overall set of beliefs. A belief is justified iff it "coherences" with the cognizer's entire set of beliefs. Insofar as holistic positive coherence theories talk about reasons at all, it makes more sense to say that a cognizer has *reason for* a belief rather than *a* reason for a belief. On a holistic theory, the reason for a belief is its coherence with the set of beliefs. In such a theory, it makes no sense to ask what reason the cognizer has for his reason.

There is some question exactly how holistic positive coherence theories are to be understood. Remember that our topic is rational cognition. If holistic positive coherence theories are to be

relevant, they must tell us something about rational cognition and epistemic norms. It seems pretty clear that a cognitive agent could not employ an epistemic norm appealing to whether a potential belief coheres with the entire set of the agent's beliefs (where coherence is some nontrivial relation that really makes reference to *all* the agent's beliefs). Such an epistemic norm would be computationally infeasible. So holistic positive coherence theories cannot plausibly be regarded as direct descriptions of epistemic norms.

A more plausible alternative is that holistic positive coherence theories are theories of warrant rather than theories of justification. Recall the distinction introduced in chapter one. Theories of epistemic justification are theories of what the agent should believe, here and now, given his current epistemic state. Theories of warrant are theories of what the agent would be justified in believing if, starting from his current epistemic state, he could complete all possible relevant cognition. Theories of justification are formulated by stating epistemic norms. Holistic positive coherence theories cannot contribute directly to that enterprise. It might be thought, however, that holistic positive coherence theories are really about warrant. They require that in order for a belief to be justified after all possible relevant cognition has been completed, it must "cohere" with the agent's overall set of beliefs. This sounds like it might well be true on some reading of "cohere", although for any specific account of coherence we would have to examine the claim more closely.

But notice that even if some holistic positive coherence theory provides a correct account of epistemic warrant, it will only do so because some other theory provides a correct account of epistemic norms, and then the theory of epistemic warrant that follows from those epistemic norms turns out to make the coherence theory true. It is important to realize that theories of epistemic warrant cannot stand alone. What makes a theory of epistemic warrant true is its relation to a theory of epistemic norms. And we have seen that the epistemic norms that might make a holistic positive coherence theory true cannot themselves appeal to a holistic positive coherence relation. So it appears that the holistic positive coherence theory really has nothing to tell us at this point about the contents of our epistemic norms.

If a coherence theory is to constitute a correct theory of epistemic justification, it must be a linear positive coherence theory. A linear positive coherence theory requires that every belief is or can be supported by linear reasons. However, such a theory cannot accommodate perception. For beliefs formed on the basis of perception, there are no other beliefs that constitute reasons for them. For instance, suppose I look at a book on my desk and judge that it is red. As the foundationalist observes, the only kind of belief that could provide a reason for thinking the book is red is a belief about my perceptual state, e.g., the belief that the book looks red. I do not normally have any such belief, but if I did that would just create further problems for the coherentist. This is because I would need a reason for the belief that the book looks red, and there are no plausible candidates for that. Perception produces a percept, and then there is a first belief adopted in response to having the percept. Whatever that belief is, whether it is about physical objects or about the percept itself, the linear positive coherence theory would require us to have a further belief that is a reason for it. But by definition, it is the *first* belief produced from the percept, so there is no further belief that could be a reason for it.

2.4 Non-Doxastic Theories

Epistemic norms cannot be externalist. Foundations theories and coherence theories exhaust the class of internalist doxastic theories, and we have seen that they cannot provide the correct norms of human rationality, so it follow that human epistemic rationality must be described by an internalist non-doxastic theory. Other theories go wrong specifically with regard to perception. Perception provides us with a percept, and from the percept we acquire a belief. The belief is normally a belief about physical objects, but occasionally it may be a belief about the percept itself. In either case, the belief is not self-justifying. If it were held on a different basis, it could

fail to be justified (even without some further belief to the effect that it should not be held). But we cannot require the cognizer to have some further belief that constitutes a reason for the perceptual belief. The whole point of perception is to be a source of new beliefs about the world—beliefs that cannot be derived from other beliefs the cognizer already holds. So the belief is not automatically justified, but it is not justified by other beliefs either. Hence it must be justified by non-beliefs. Having gotten this far, it is obvious that what really justifies the belief is the percept itself. In other words, our procedural knoweldge for how to cognize contains epistemic norms licensing the adoption of beliefs directly on the basis of having percepts, and without any further beliefs about the percepts themselves.

I think that this would be regarded as completely commensensical if philosophers were not so enamoured of the doxastic assumption. Recall the argument given above for the doxastic assumption. Procedural epistemic justification is supposed to be concerned with what to believe. But in deciding what to believe, we can only take account of something insofar as we have a belief about it. Thus only beliefs can be relevant to what we are justified in believing. First notice that this argument for the doxastic assumption could not possibly be right, because it is self-defeating. If this argument were right, we could only take account of our beliefs insofar as we have beliefs about our beliefs, and then an infinite regress would loom. There has to be something about beliefs that makes them the sort of thing we can take account of without having beliefs about them. What could this be?

What is it to take account of something in the course of cognition? It is to use it in our cognitive deliberations. We can take account of anything by having a belief about it, but cognition has to start somewhere, with things that we don't have beliefs about. Obviously, it can start with beliefs. The reason it can start with beliefs is that they are internal states, and cognition is an internal process that can access internal states directly. Cognition works by noting that we have certain beliefs and using that to trigger the formation of further beliefs. However, it is *cognition* that must note that we have certain beliefs—*we* do not have to note it ourselves. The sense in which cognition notes it is metaphorical—it is the same as the sense in which a computer program accessing a database might be described as noting that some particular item is contained in it.

Epistemologists have a lamentable tendency to over-intellectualize cognition. Human beings are *cognitive machines*. We are unusual machines in that our machinery can turn upon itself and enable us to direct many of our own internal operations. Many of these operations, like reasoning, can proceed mechanically, without any deliberate direction or intervention from us, but when we take a mind to we can directly affect their course. For example, we can, at least to some extent, decide what to think about, decide not to pursue a certain line of investigation, and to pursue another one instead. There must, however, be a limit to the extent to which we are *required* to do this. After all, the processes by which we do it are a subspecies of the very processes in which we are intervening. If we had to explicitly direct all of our cognitive processes, we would also have to direct the ones involved in doing the directing, and we would again have an infinite regress.

The significance of this is that we don't *have* to think about our cognition in order to cognize. It is important, for various reasons, that we *can* think about it when the need arises, but we don't have to and don't usually do it. Thus cognition can proceed by moving from beliefs to beliefs without our thinking about either cognition or the beliefs. By virtue of doing the cognizing we are thinking about whatever the beliefs are about, not about the beliefs themselves. This explains the sense in which cognition can take account of our having certain beliefs without our having to have beliefs to the effect that we have those beliefs. But note that it explains much more. In precisely the same sense, cognition can take account of other internal states, for example, percepts, without our having to have beliefs to the effect that we are in those states. Thus there is no reason, in principle, why cognition cannot move directly from percepts to

beliefs about the physical objects putatively represented by the percepts.

Cognition can make use of any states to which it has direct access, but those are just the internal states. So cognition can make use of any internal states without our having beliefs about those states, and correspondingly our epistemic norms can appeal to any internal states—not just beliefs. Such nondoxastic norms only seemed puzzling because we were implicitly assuming the intellectualist model of the way epistemic norms regulate belief. Given the way epistemic norms actually operate, all that is required is that the input states be directly accessible. Belief states are directly accessible, but so are a variety of nondoxastic states like perceptual states and memory states. Thus there is no reason why epistemic norms cannot appeal to those states, and the rejection of the doxastic assumption and the move to direct realism ceases to be puzzling.

Once the intellectualist model of procedural norms is rejected, we can see that epistemic norms can in principle appeal to any internal states of the cognizer—not just to beliefs. Accordingly, there is no reason we cannot have epistemic norms licensing a cognitive move directly from percepts to beliefs about physical objects. The resulting theory looks much like tradtional foundationalism, differing only in that the foundations consist of percepts rather than beliefs about percepts. A theory of this sort is what is called "direct realism". The terminilogy is not very perspicuous, but what it is intended to capture is the idea that the move from percepts to beliefs about the real world is direct and unmediated by beliefs about the agent's internal states. The epistemological theory developed in this book will be a version of direct realism.

To build a direct realist theory, we must articulate the epistemic norms governing both the formation of perceptual beliefs on the basis of percepts, and the derivation of other beliefs on the basis of inference. Before we can do that, we must say much more about inference itself. The next section will provide a skeletal account of certain aspects of inference, but the bulk of the theory will be presented in chapter three.

3. Epistemic Cognition in Context

Epistemologists have generally studied epistemic cognition in isolation. This is responsible for the common idea that a cognitive architecture is to be evaluated somehow in terms of its tendency to produce true beliefs. If one ignores the fact that epistemic cognition is just part of a large cognitive system aimed at producing actions, then there seems to be no way to evaluate the system of epistemic cognition except in terms of some abstract property like truth. However, the architecture for epistemic cognition is just part of an overall cognitive architecture, as described by the doxastic-conative loop. As such, an architecture for epistemic cognition should be evaluated instrumentally, in terms of how well it helps the overall cognitive architecture in which it is embedded to achieve its design goal. In order to perform such an instrumental evaluation, we must see more clearly just how epistemic cognition is integrated into the larger cognitive architecture.

3.1 Practical Cognition

Let us focus first on practical cognition. We can think of practical cognition as having three parts, as diagrammed in figure 2.2. First, the agent evaluates the world (as represented by its beliefs) and forms goals for changing various aspects of the world to make it more to its liking. Then the agent constructs plans for how to achieve its goals. Finally, it executes those plans, producing actions.

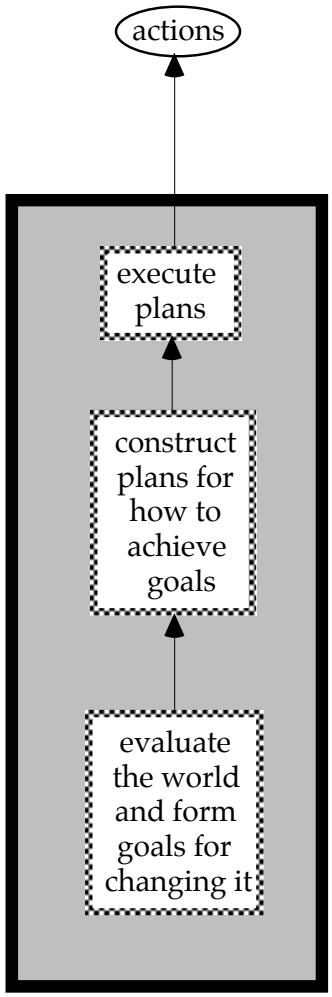

Figure 2.2 Practical cognition

The point of epistemic cognition is to provide the information required for practical cognition. This information is encoded in the form of beliefs, and the beliefs are used by all three modules of practical cognition, as diagrammed in figure 2.3. The agent evaluates the world, as represented by its beliefs, in order to form goals for changing various aspects of the world to make it more to its liking. So goals are adopted on the basis of the agent's beliefs about the way the world is. Plans are constructed on the basis of beliefs about how the world works. That is, in deciding how to act the agent must employ beliefs about what the consequences of its actions are apt to be. And once a plan of action is adopted, its execution will typically turn upon the agents beliefs about the world in at least two ways. First, the timing of actions may depend upon when various events occur, and the agent must have beliefs about that. Second, the agent will typically have to monitor the execution of the plan to make sure that various actions have the envisaged results.

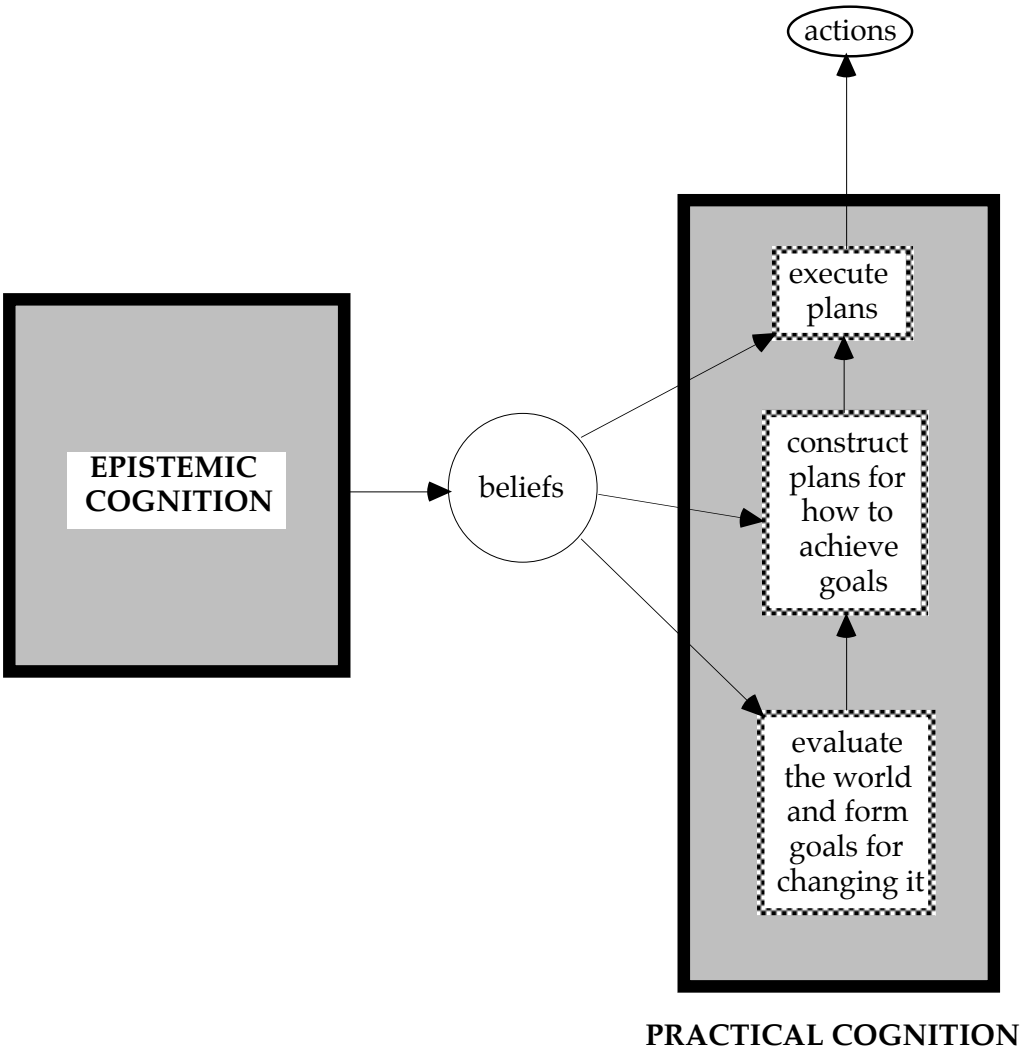

Figure 2.3 The subservient role of epistemic cognition

3.2 Ultimate Epistemic Interests

The basic source of information for epistemic cognition is provided by perception. We can imagine a cognitive architecture in which the agent just reasoned at random from its perceptual input, and when it happened upon conclusions that were of use to practical cognition, they would be used appropriately. But such a system would be grossly inefficient, and it is clear that that is not the way human cognition works. Any practical system of epistemic cognition must take account of what kind of information would be useful in the agent's practical endeavors, and focus its epistemic efforts accordingly. Practical cognition poses queries which are passed to epistemic cognition, and then epistemic cognition tries to answer them. Different queries are passed to epistemic cognition depending upon what practical cognition has already occurred. For example, once the agent has adopted a particular goal, it tries to construct a plan for achieving it. In order to construct such a plan, a query should be passed to epistemic cognition concerning what actions are apt to achieve the goal and under what circumstances. Similarly, when the agent adopts a plan, the timing of the execution will depend upon when various things happen in the world, so practical cognition should send a corresponding query to epistemic cognition. Epistemic cognition answers these queries by producing appropriate beliefs, which are then used by practical cognition. The queries posed by practical cognition comprise the set of *ultimate*

epistemic interests. Thus we can expand the diagram of figure 2.3 to produce figure 2.4.

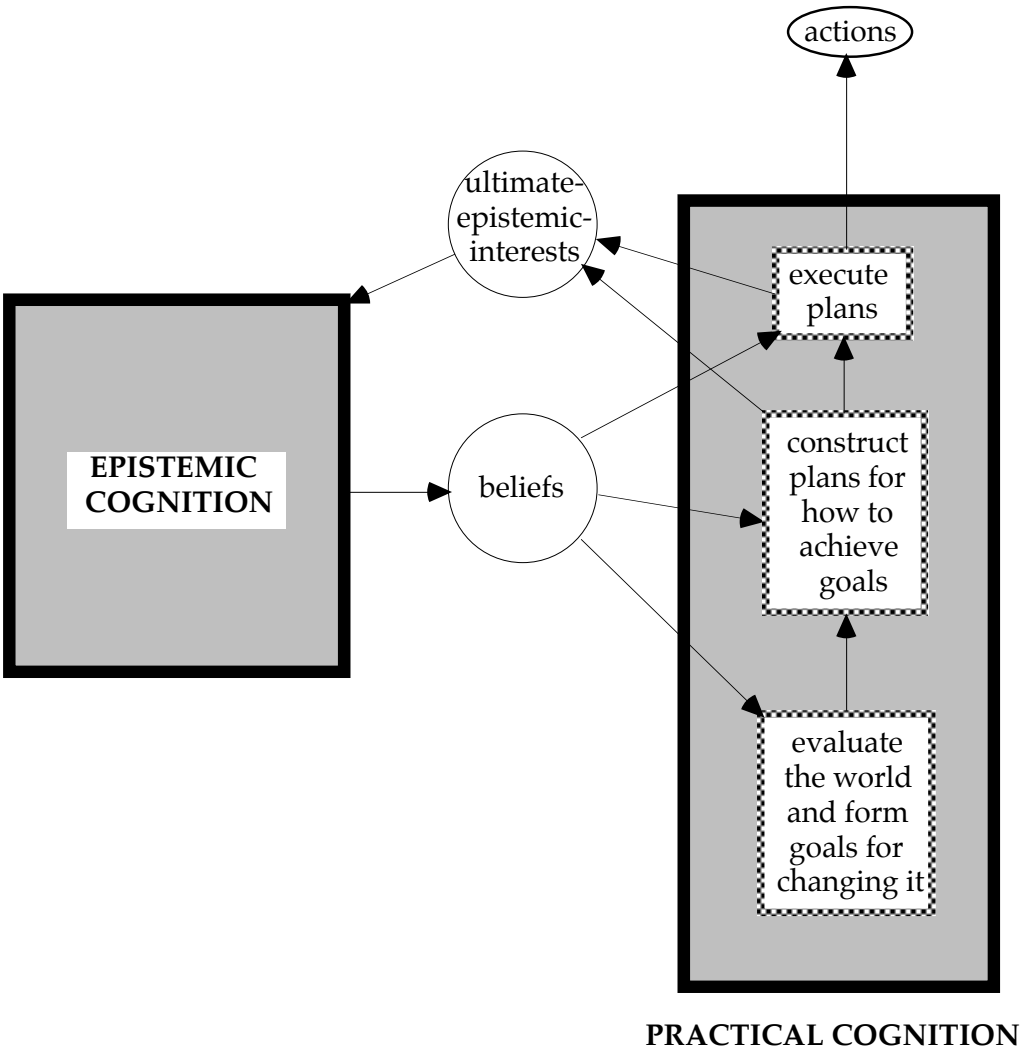

Figure 2.4 Query passing

3.3 Bidirectional Reasoning

Apparently the course of epistemic cognition must be driven by two different kinds of inputs. New information is input by perception, and queries are passed to it from practical cognition. This produced bidirectional reasoning. The agent reasons forward from perceptual input, but such reasoning is not directed in any way by the queries. To make use of the queries in its reasoning, the agent also reasons backward. The queries are *epistemic interests*. They represent things the agent wants to know. The agent may be equipped with inference-schemes that would allow it to answer a particular query if it knew somethng else. For example, given an interest in knowing (P & Q), the agent could satisfy that interest if it knew P and Q separately. So the agent *reasons backward* and adopts interest in P and Q. These are derived epistemic interests. In this way the agent can reason backward from its ultimate epistemic interests and forwards from perceptual input until it finds itself in a situation in which forward reasoning has produced a belief that answers a query at which it has arrived by backward reasoning. This *discharges* the interest, enabling the agent to use the answer to that query in answering the query from which it was derived, and so on. By interleaving backward and forward reasoning, the

agent is thus able to use its ultimate epistemic interests to help guide its epistemic cognition and make it more efficient in providing the information needed for practical cognition.

We saw above that epistemic cognition combines epistemic reasoning with the operation of Q&I modules. The preceding remarks apply to reasoning. Q&I modules proceed more mechanically, producing an automatic output from various inputs. The inputs are are in the form of new beliefs produced by epistemic cognition. We can accordingly expand the diagram of figure 2.4 by including some of the structure of epistemic cognition, as in figure 2.5. This represents the basic interface between epistemic and practical cognition.

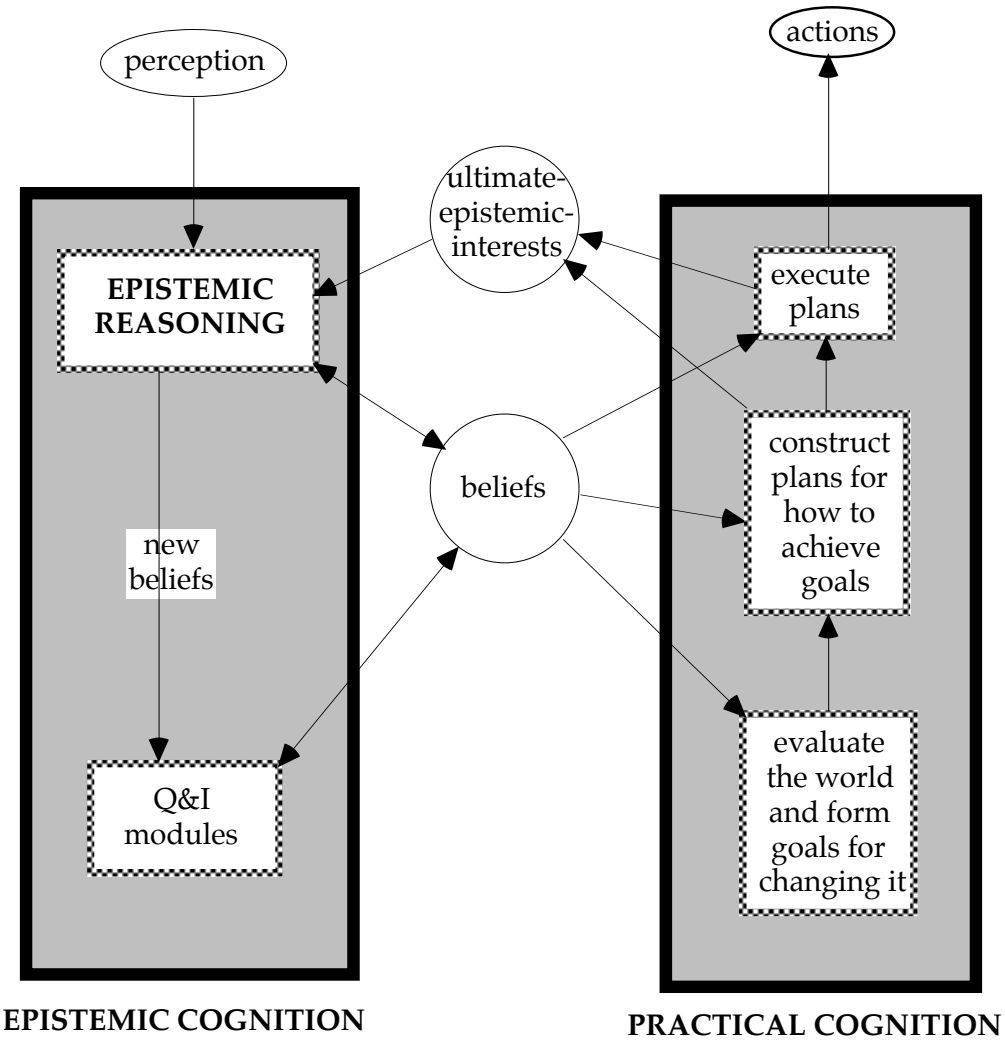

Figure 2.5 The basic interface.

3.4 Empirical Investigation

The architecture diagrammed in figure 2.5 has the agent attempting to answer queries posted by practical cognition just by thinking about them. This is the kind of cognition that philosophers have traditionally focused upon. But many questions of practical interest cannot be answered just by thinking about what the cognizer already knows. To answer even so simple a question as "What time is it?", the agent may have to examine the world—at least look at a clock. More difficult questions may require looking things up in reference books, talking to other cognizers, searching one's closet, performing experiments in particle physics, etc. These

are *empirical investigations*. What is characteristic of empirical investigations is that epistemic interests give rise to actions aimed at acquiring the information of interest. Actions are driven by practical cognition, so this involves a connection whereby epistemic cognition initiates further practical cognition. Practical cognition begins with the formation of goals, and then looks for plans that will achieve the goals. Accordingly, the mechanism whereby epistemic cognition can initiate practical cognition is by introducing "epistemic goals"—goals for the acquisition of information.

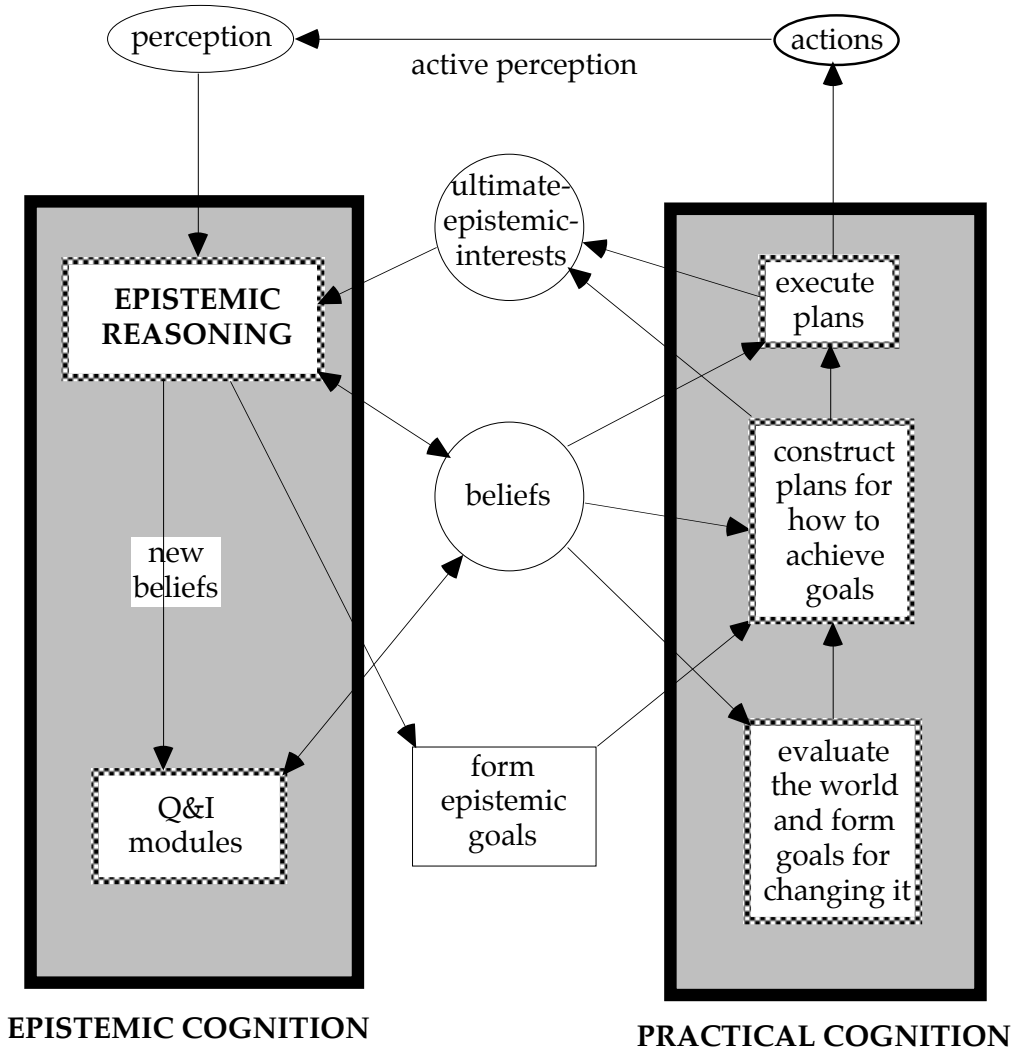

Figure 2.6 Empirical investigation

Empirical investigation introduces another interconnection between epistemic and practical cognition. The architecture diagrammed in figure 2.5 treats perception as spontaneous and undirected. However, there is a familiar distinction between looking and seeing. Seeing is passive, but looking involves putting yourself in an appropriate situation for seeing what you want to see. The same distinction applies to other forms of perception as well. We distinguish between sniffing and smelling, touching and feeling, listening and hearing, etc. In general, there is a distinction between "active" and "passive" perception, where active perception consists of putting oneself in an appropriate situation for passive perception. Some of the actions generated by practical cognition in the course of an empirical investigation will typically involve active

perception. For instance, the agent may looking something up in a reference book by turning to the appropriate page and then looking for the information on the page.

Empirical investigation is accommodated by adding two links to the architecture of figure 2.5, generating that of 2.6. One link is from epistemic cognition, via the formation of epistemic goals, to practical cognition, and the other links is from practical cognition to epistemic cognition via active perception. Adding these links has the consequence that we get loops from practical cognition to epistemic cognition, then back to practical cognition, and so on. Practical interests give rise to epistemic interests, which may in turn produce epistemic goals, which may initiate a search for a plan for how to find the desired information, which passes a new query to that effect back to epistemic cognition. That may give rise to new epistemic goals and further loops back through practical cognition.

3.5 *Reflexive Cognition*

Although epistemic cognition is initiated by practical cognition, it need not be *directed* by practical cognition about how to answer questions. That would lead to an infinite regress, because practical cognition always requires beliefs about the world. If we could not acquire some such beliefs without first engaging in practical cognition, we could never get started. This indicates that there must be a *default control structure* governing epistemic cognition, and in particular, governing the way in which it tries to answer questions. However, in human beings it is also possible to override the default control structure. For example, consider taking an exam. A good strategy is often to do the easy problems first. This involves engaging in practical cognition about how to arrange our epistemic tasks. Our default control structure might, for example, take them in the order they are presented, but we can think about them and decide that it would be better to address them in a different order. Cognition about cognition is *reflexive cognition*. We can distinguish between two kinds of reflexive cognition:

3.5.1 Practical Cognition about Cognition

When we redirect the course of our cognition by thinking about it and deciding what it would be best to think about first, we are engaging in practical cognition about how to cognize. Clearly it is a good idea for a sophisticated rational agent to be able to modify the course of its own epistemic endeavors by engaging in practical cognition about how best to pursue them. An agent can learn that certain natural (default) strategies are unlikely to be effective in specific circumstances, and new strategies may be discovered that are more effective. The latter are often obtained by analogy from previous problem solving. It is desirable for a rational agent to use practical cognition in this manner to direct the course of either epistemic or practical cognition.

Allowing practical reasoning to intervene in the course of epistemic cognition can also be useful in keeping the cognizer out of trouble. The gambler's fallacy provides an illustration of this phenomenon. Contrary to popular opinion, reasoning in accordance with the gambler's fallacy is not automatically irrational. Such reasoning proceeds in terms of a perfectly good defeasible reason of the form "Most A's are B's, and this is an A, so (defeasibly) this is a B". In the case of the gambler's fallacy, there is a defeater for this reason, but it is surprisingly difficult to say precisely what the defeater is.[3] Most people who consciously refrain from reasoning in accordance with the gambler's fallacy do so not because they understand precisely what is wrong with it, but because they know from experience that they get into trouble when they reason that way. It is practical reasoning that intervenes to prevent the epistemic reasoning.

It is interesting to consider just how much power an agent should have over its own cognition. It is preumably undesirable for an agent to be able make itself belief something just

[3] For a detailed discussion of the reasoning involved in the gambler's fallacy and the defeater responsible for making the reasoning ultimately incorrect, see Chapter Nine of my [1989].

because it wants to believe it. Belief should be more constrained by specifically epistemic cognition. On the other hand, it seems to be a good thing for an agent to be able to determine what it is going to think about next. But even here, there are dangers. Every scientist or scholar has experienced the temptation to ignore possible possible difficulties for his or her theories.

There are limits to how much power reflexive cognition should have, but it is not clear how the limits should be set. That is a topic the needs more research.

We can accommodate practical cognition about cognition by adding one link from practical cognition to epistemic cognition, as in figure 2.7.

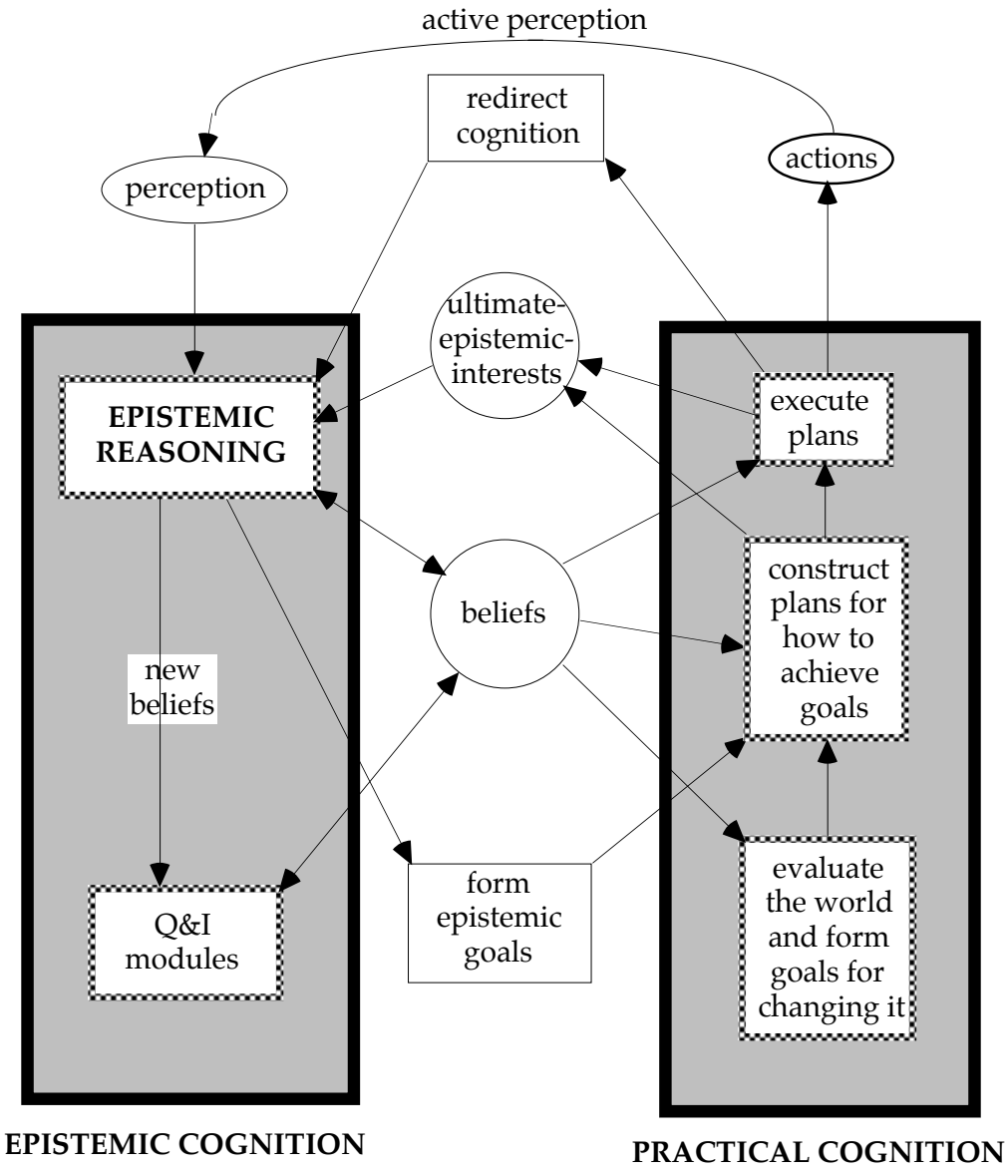

Figure 2.7 Epistemic cognition in context

3.5.2 Epistemic Cognition about Cognition

We also engage in epistemic cognition about cognition. That is an automatic consequence of our engaging in practical cognition about cognition, because practical cognition always presupposes prior beliefs, produced by epistemic cognition, about whatever the practical cognition

is directed at. Thus we must be able to form beliefs about what our epistemic interests are, what reasoning we have done or are in the course of doing, what epistemic strategies have been useful on similar tasks in the past, etc. Beliefs about various aspects of our current cognition are produced by introspection. Functionally, introspection is a form of perception. From an implementational perspective, introspection is by far the simplest form of perception. Unlike vision or hearing, we do not need an "organ of introspection". Introspection can be implemented by just having some of the information the cognitive system is using in its processing also produce beliefs. This is a simple matter of programming or wiring. Note that introspection, so conceived, has an obvious and important functional role in cognition.

Introspection produces beliefs about various mental states and mental processes. Introspection might inform me, for example, that I am reasoning, but it will generally be more useful if it can also inform me *how* I am reasoning, i.e., what conclusions I am drawing from what earlier conclusions. This involves introspecting the content of a mental state. Mental states having such contents are called "propositional attitudes". In order to form the belief that a certain mental state token has a particular content, I must be able to think about both the token and the attributed content.

Thinking about something involves having a mental representation that designates it. In visual perception, percepts provide us with a way of thinking about the objects we perceive, and so play a representational role in our thoughts. We need something similar in introspection. When we introspect a mental occurrence, introspection itself produces a mental representation of the introspected occurrence which we can then employ in thinking about the occurrence.

But notice that if my thought is that a certain mental state token has a particular content, I must also be able to think about the content. "Thinking a thought" consists of "thinking its content", i.e., *using* the content in a certain way in cognition. But thinking *about* the content requires the use of a mental representation that designates the content. We have a use/mention distinction here. What this shows is that reflexive epistemic cognition involves the use of a representational device which, when applied to a mental content, produces a mental representation of the content. Given a content p, let «p» be the corresponding mental representation of it. These are like mental quotation names for mental contents.

It is generally most useful to think about the contents of thoughts when we are also in a position to judge whether the content is true or false. For example, one important function of reflexive epistemic cognition is to discover that certain kinds of non-deductive reasoning are unreliable under specific circumstances. For example, I may discover that color vision is unreliable when the objects seen are illuminated by colored lights. Then if I later know that I am seeing somethng illuminated by colored lights, that would defeat an inference to the conclusion that it is the color it appears.

Discovering that a certain kind of reasoning is unreliable under specific circumstances amounts to discovering that it often produces false conclusions under those circumstances. To know that, we must be able to judge whether a particular content is true or false. How can we do that? The obvious proposal is to adopt an analogue of Tarski's T-schema:[4]

> From p, infer ⌜«p» is true⌝.
> From ⌜«p» is true⌝, infer p.

These reasoning-schemas seem obvious, but there is a surprising problem associated with them. If an agent equipped with these reasoning-schemas is also able to form self-referential beliefs, these reasoning schemas will lead directly to the liar paradox. Consider, for instance, a philosopher who knows that he is always in a state of befuddlement until he has his first cup of coffee in the

[4] Tarski [1956].

morning. Awakening with this realization, he may think to himself, "The first philosophical thought I have this morning will be false". After he has his first cup of coffee, he realizes that that *was* his first philosophical thought of the morning. Employing the T-schema, he infers that if his thought was true then it was false, and if it was false then it was true. It follows by simple logic that it was both true and false. So the T-schema leads to a contradiction.

There is a vast literature in philosophical logic on the liar paradox.[5] To a large extent, solutions to the liar paradox proceed by proposing restrictions to the T-schema which preclude its use in paradoxical instances. A noteworthy feature of all such solutions is that they preclude our attributing truth to some mental contents which are themselves unproblematic and to which it seems, intuitively, that we should be able to attribute truth. This at least strongly suggests that those accounts, even if they do successfully avoid the paradox, are not correct descriptions of the human epistemic norms governing reasoning about truth.

A perplexing possibility arises here. Perhaps the T-schema really is a correct descriptions of our built-in epistemic norms, despite the fact that it is inconsistent. Then what? I am tempted by a radical solution to the liar paradox. I suspect that, despite the inconsistency, the T-schema is an accurate description of our rules for reasoning about truth. Any attempt to avoid the liar paradox by restricting the T-schema seems *ad hoc* and unmotivated. My suggestion is that the solution to building an introspective reasoner that that does not collapse in the face of the liar paradox lies in damage control rather than damage avoidance. When a reflexive reasoner gets into trouble, it is able to apply practical cognition to the situation and back out gracefully without necessarily solving the problem that led to the difficulty. This is a general phenomenon of considerable importance. It is illustrated by a number of situations. For instance, mathematicians tend to be very cautious about accepting the results of compicated proofs. But the proof, if correct, is simply an exercise in deductive reasoning. How is such caution possible? Why doesn't their rational architecture force them to just automatically accept the conclusion? The answer seems to be that they have learned from experience that complicated deductive reasoning is error prone. Practical cognition about epistemic cognition has the power to intervene and prevent acceptance of the conclusion until the proof is adequately checked. That practical cognition has this power is just a brute fact about our rational architecture.

A related phenomenon is common in any intellectual discipline. An investigator formulating a theory will frequently encounter difficulties that he cannot answer in the early stages of the investigation, but this does not automatically move him to give up his investigation. He may suspect that there is a way for the theory to handle the difficulty even if he cannot initially see what it is. This can be a perfectly reasonable attitude based upon the observation that the theory has been successful in avoiding many other difficulties whose solutions were not intially obvious. If the investigator were constrained to blindly follow the default rules of epistemic cognition, he would have to give up the theory *until* he is able to resolve the difficulty, but because he anticipates that he will eventually be able to resolve it, he instead retains the theory and continues to look for a solution to the problem while developing the theory in other ways. This example is particularly interesting because the investigator may literally have inconsistent beliefs but postpone giving up any of them. Instead of rejecting the beliefs, he retains them but does not draw any conclusions from the resulting contradiction. Practical cognition imposes a kind of damage control that encapsulates the problem until a solution is found.

My suggestion is that precisely the same phenomenon may be involved in the liar paradox. Reasoning about the liar sentence gets us into cognitive difficulties. We do not immediately see our way out of them, but we do not simply go crazy as a result. Instead, practical cognition intervenes to encapsulate the problem until we find a solution. The difference between this case

---

[5] For a sampling of the literature on the liar paradox, see Martin [1970, 1984]. For an extremely interesting proposal from AI, see Perlis [1985].

and the previous case is that there may be no solution to the liar paradox. If the T-schema is an accurate description of part of our rational architecture, then our rational architecture is logically inconsistent, and there is nothing we can do about it. We cannot change the rules for reasoning about truth, because given a conceptual role theory of concept individuation, those rules (even if they are inconsistent) are *definitive* of the concept of truth. However, for a reflexive reasoner, this need not be catastrophic.

4. Conclusions

This chapter has defended three main theses. The first is that epistemic cognition consists of a combination of reasoning and Q&I modules, with reasoning playing the dominant role but Q&I modules doing a lot of the work of day-to-day epistemic cognition. The second is that the epistemic norms governing human epistemic cognition constitute a kind of direct realism. The third is that for many purposes, epistemic cognition cannot be studied in isolation but must instead be viewed in a larger context that includes its interrelations with practical cognition.

Because Q&I modules are introspective black boxes, their operation does not fall within the purview of rationality, so this book will have little to say about specific Q&I modules. The interaction of reasoning and Q&I modules is introspectible (at least in humans), so there will be some further discussion of how that is handled in an architecture for rational cognition.

Turning to direct realism, our conclusion thus far is only that some form of direct realism is true. This is a very schematic conclusion. What remains is to complete the account by giving precise descriptions of the epistemic norms comprising our system of epistemic cognition. That is a large task. Epistemologists have often sought some kind of unifying characterization that would generate all correct epistemic norms, without our having to discuss each one separately. Such a unifying theory would be very nice if it were available, but it is extremely doubtful that such a theory is possible. This is a consequence of the nature of procedural norms. Such norms instruct us to do various things under various circumstances and prohibit us from doing other things. These norms have to be rather specific because, as we saw above, they must take as input only features of the present circumstances that are directly accessible to our cognitive systems. This precludes the possibility of the norms appealing to sweeping general features of the circumstances (features such as the belief being produced by a reliable process). Compare the norms for bicycle riding. These are going to be very specific, including such things as, "If you feel yourself losing momentum then push harder on the pedal" and "If you think you are falling to the right then turn the handlebars to the right". Epistemic norms will be equally specific, telling us things like as "If something looks red to you and you have no reason for thinking it is not red then you are permitted to believe it is red". There is no more reason to think that we can combine all epistemic norms into one simple general formula than there is for thinking there is a single simple formula governing the use of the pedals, the handlebars, the brakes, and so on, in bicycle riding. Procedural norms cannot work that way.

Most of our epistemic norms govern reasoning. The simplest way to formulate them is within a general framework of reason-schemas and reasoning. The next chapter will undertake the development of such a general framework. Subsequent chapers will explore the epistemic norms governing specific kinds of reasoning, e.g., reasoning from perception, inductive reasoning, probabilistic reasoning, and so forth. In developing the general framework, we will make essential use of the interactions between epistemic and practical cognition that were noted above.