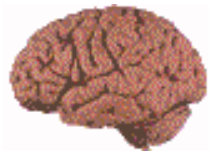

RATIONAL COGNITION

John L. Pollock

Department of Philosophy
University of Arizona
Tucson, Arizona 85721
(e-mail: pollock@arizona.edu)

Table of Contents

1
OVERVIEW

1. Epistemic and Practical Cognition

Perhaps the most striking feature of human beings is that they think, and when they act their actions are directed by their thoughts. In short, they are *cognitive agents*. They engage in varous kinds of cognition, and their interactions with the world are mediated by that cognition. Some aspects of human cognition are rather mechanical, and not noticeably different from cognition that occurs in lower animals. An example would a reflex response to pain. But other aspects of human cognition are very sophisticated, including forming complicated beliefs and expectations about the world, adopting goals that lower animals could not even conceive of, and constructing complex plans for how to act in an attempt to achieve those goals. Some of these more sophisticated forms of cognition are called "rational". Cognitive agents engaging in rational cognition can exhibit greater or lesser degrees of rationality, and we can assess their cognitive behavior in those terms. Human beings are not totally rational, but it is plausible to think that they are for the most part rational. That is, although they do sometimes behave irrationally, that is sufficiently unusual to be noteworthy.

Granted that humans are largely rational, what is it to be rational? What is rational cognition, and what are the dimensions along which some cognition can be more rational than other? The objective of this investigation is to produce a general theory of rational cognition.

It is customary to distinguish between two kinds of rational cognition — epistemic cognition and practical cognition. Epistemic cognition is cognition regarding what to believe, and practical cognition is cognition regarding what to do. Both kinds of cognition are essential to sophisticated interaction with the world, and the degree to which an agent engaging in either kind of cognition is behaving rationally can vary. One of the most fundamental questions in the theory of rational cognition is how epistemic and practical cognition are related.

A natural supposition would be that epistemic and practical cognition differ only in what they are about — not in the cognitive processes themselves. On this view, there is just one kind of rational cognition, and sometimes it is directed at what to believe and other times at what to do. This has been a common view in AI. It is important to realize that this view is wrong. There are intimate connections between these two kinds of cognition, but for a variety of reasons they must be distinguished from each other.

One simple way of trying to unify epistemic and practical cognition would be to suppose adopting beliefs is one kind of thing we "do", and so epistemic cognition is just practical cognition about what to believe. Such an assimilation of epistemic cognition to practical cognition is bound to fail, however, because in deciding what to do we must take account of contingent features of the situation in which we find ourselves. Our access to these contingent features is through epistemic cognition. Thus practical cognition presupposes factual beliefs produced by epistemic cognition. Any attempt to reduce epistemic cognition to practical cognition would thus lead to an infinite regress.

One might be tempted by the opposite reduction, identifying cognition about what to do with cognition about what we *should do*. The latter at least has the form of a factual question of the sort falling within the purview of epistemic cognition. One of the main theses of this book will be that, to a large extent, this reduction works. But it cannot entirely reduce practical

cognition to epistemic cognition, because given the conclusion that I should do something, some further kind of practical decision is required to get me to actually do it. In other words, epistemic cognition can issue in beliefs about what to do, but only practical cognition can issue in the actions themselves.

If the reader is still unconvinced by these considerations, let me point out a rather deep logical difference between epistemic and practical cognition. Suppose you have equally good reasons for believing something and for disbelieving it. For instance, Jones, whom you trust, tells you that it is raining. This makes it defeasibly reasonable to believe that it is raining. But then Smith, whom you trust equally, tells you that it is not raining. What should you believe? Clearly, in the absence of further evidence, you should withhold belief, remaining agnostic on the question of whether it is raining. This is sometimes put by saying that epistemic cognition is *skeptical*. Now contrast this with practical cognition. Consider the medieval tale of Buridan's ass, who starved to death midway between two equally succulent bails of hay because he could not decide from which to eat. Clearly, this is irrational. When faced with a practical choice between two equally desirable actions not both of which can be performed, a rational person will choose between them randomly rather than being frozen into inaction. This is sometimes put by saying that practical cognition is *credulous*.[1] This fundamental difference between epistemic and practical cognition makes it reasonably clear that they are not instances of a single kind of cognition differing only in what it is cognition about.

Having argued that epistemic and rational cognition are genuinely different kinds of cognition, it may be wondered whether they are totally distinct. One of the main theses of this book will be that that they are not. The division is not a clean one. Epistemic cognition is cognition aimed at forming and maintaining our beliefs, and practical cognition is cognition aimed at acting, but there may be large parts of cognition that play important roles in serving both ends. In fact, the division is largely a matter of degree, turning upon how directly an aspect of cognition is connected with the goal of either forming beliefs or taking action. If we are sufficiently generous in drawing connections, all of epistemic cognition will become included in practical cognition, because presumably the whole point of forming beliefs is to acquire information of use in deciding how to act. And conversely it will be argued that even aspects of cognition that seem very directly connected with action, such as cognition aimed at constructing plans for achieving goals, are performed by the mechanisms of epistemic cognition.

It should be noted that not all aspects of cognition are included within *rational* cognition. Even some very high level cognition, like visual processing, escapes the purview of rationality. A person who sees pink elephants may be sick or demented or inebriated but not thereby irrational. One question that this investigation will try to answer is what makes some cognition rational and other cognition non-rational, and why we even make such a distinction. In studying rational cognition we are studying information processing and decision making *processes*. Calling them "rational" is to assess them in some way. Presumably the nonrational cognitive processes are those that can be evaluated in the appropriate manner. Theories of rationality are, in part, theories about the nature of this assessment.

2. Studying Rationality

Philosophers have traditionally studied rationality by employing their "philosophical intuitions" about whether various kinds of cognitive behavior are rational or irrational, and

---

[1] The skeptical/credulous distinction is due to Horty, Touretsky, and Thomason.

propounding theories that attempt to capture those intuitions. This has led to some notable successes. For example, decision theory is ultimately defended on such grounds. If you doubt this, reflect upon the fact that the central tenet of decision theory is that we should try to maximize expected utility. This is importantly different from saying that we should try to maximize actual utility. It tells us how to mix considerations of utility and probability in choosing actions. But consider, why should we try to maximize expected utility? The untutored individual does not find this thesis immediately obvious when he or she first encounters it. At least, *I* didn't as an undergraduate. But one can become convinced of it by reflecting upon actual examples of practical deliberation. In doing this, one is relying upon some kind of pre-analytic intuition of what is rational and what is not.

As successful as this philosophical methodology might be, one cannot help but wonder about its credentials. What are these so-called "philosophical intutions", and how is it that they are able to inform us about rationality? Even if we grant that they are unproblematic and that they can reveal the structure of rationality, a deep question remains. *Why* should rational cognition have the structure thus revealed? What would be wrong with a cognitive agent that worked differently? These are fundamental questions that must be addressed by a theory of rationality.

It will be argued that considerable light can be thrown upon these questions by approaching rationality from a different direction. This is to approach it from the "design stance", to use Dennett's [1987] felicitous term. We think about how to build a cognitive agent, and then try to relate rationality to the assessment of such an agent. To do this we must give some thought to what the objectives are in building a cognitive agent. This involves thinking of rational cognition as the solution to a design problem. We must consider what that design problem is, and how it might conceivably be solved. I will argue that quite general features of this design problem suffice to generate much of the structure of rationality. General logical and feasibility constraints have the consequence that there is often only one obvious way of solving certain design problems, and that is the course taken by human rationality.

3. Epistemology

Epistemology is the discipline that studies rational epistemic cognition. Epistemologists have traditionally propounded theories of how, rationally, to proceed in pursuing various epistemic endeavors. The results have been theories of our knowledge of the material world, our knowledge of other minds, induction, scientific theory confirmation, etc.

The central concept of traditional epistemology was that of a justified belief. Epistemologists propound theories of what we should or should not believe under various circumstances, where the "should" is a uniquely epistemic "should". Justified beliefs are those it is "epistemically permissible" to hold.

I would like to identify justified beliefs with those held consonant with the dicates of epistemic rationality. However, the situation is a bit more complicated than that. The difficulty is that "epistemic justification" is a term of art, not a term of ordinary discourse, and in recent years it has become apparent that there is more than one important concept that might reasonably be called "epistemic justification". It seems clear that these different concepts have often been confused with one another in the epistemological literature.

Epistemology was traditionally motivated by the question, "How do you know?" At first blush, this would seem to be about knowledge, but it is so only indirectly. It is about *how* we know. This makes it a question about the rational procedures that give rise to our knowledge of the world—rational procedures for belief formation. In arriving at beliefs, epistemic agents

follow various procedures. Some of these procedures are epistemically praiseworthy, and others are epistemically blameworthy. There is a procedural sense of epistemic justification according to which a belief is epistemically justified iff it was arrived at or held on the basis of procedures that are epistemically praiseworthy. This is the traditional concept of epistemic justification—the concept that concerned Descartes, Hume, Kant, and most of their descendants. *Procedural epistemology* is the study of this concept of epistemic justification.[2]

In 1963, Edmund Gettier published his famous paper on the analysis of "S knows that P", and epistemology has never been the same since.[3] The Gettier problem is a seductive one, and a voluminous amount of philosophical ink has been spilled on it.[4] Many young epistemologists have been raised on the idea that this is the central problem of epistemology. In their attempts to solve the Gettier problem, many epistemologists have focused on a concept of epistemic justification that is something like "What turns true belief into knowledge". But one of the lessons to be drawn from the massive literature on the Gettier problem is just what a perverse concept knowledge is. What the Gettier problem shows is that a person can have a true belief, behave with complete epistemic and rational propriety, and still fail to have knowledge through no fault of their own. What is happening here is indeed an interesting problem, but note that epistemology was originally driven by concern with the notions of epistemic and rational propriety. That is what procedural justification concerns. What the Gettier problem shows is that knowledge is less directly connected with procedural justification than we thought.

A helpful way of viewing the difference between these two concepts of epistemic justification is from the design stance. Viewed from this perspective, we can roughly divide the interests of epistemology into *procedural epistemology* and *descriptive epistemology*. Procedural epistemology is directed at how to build the system of cognition, whereas descriptive epistemology concerns how to describe what the system is doing once it is running. Thus, for example, rules for reasoning become part of procedural epistemology, but the analysis of "*S* knows that *P*" is assigned to descriptive epistemology.

4. Practical Rationality

Practical rationality is concerned with those features of rational cognition that pertain most directly to the selection of actions to be performed. There is no simple label for this subdiscipline of philosophy. Philosophers often talk about practical reasoning, but that may not come to quite the same thing. This is for two reasons. First, there would seem to be more to practical cognition than reasoning. An essential part of practical cognition concerns the performance of actions. In practical cognition we do ot just reason *about* actions—we do them. That seems to involve something over and above reasoning.

A second reason for being careful about assimilating practical cognition to the philosopher's practical reasoning is that many philosophers associate practical reasoning with ethics in a strong way. For them, practical reasoning about what to do *includes* moral reasoning. However, the researchers investigating practical cognition would not want to build in moral reasoning as a mandatory part of what they are studying. The dominant view of practical cognition is that it is cognition about how to act in one's own self-interest.

On the other hand, the relationship between practical cognition and morality is an interesting

---

[2] This is persuasively documented by Mark Kaplan [1985].

[3] Gettier [1963].

[4] Shope [1983] contains a good collection of essays on the Gettier problem.

one, well worth pursuing. Practical cognition is defined generally as cognition about what to do. There is no requirement in the definition that such decisions are based entirely on the agent's self-interest. It is customary to study cognitive agents in isolation, in which case there is nothing else practical decisions could appeal to. However, for agents embedded in a social structure, the question arises how to get them to cooperate. They are faced with coordination problems, and the imposition of moral constraints is presumably a solution. But it is a very interesting question how to build morality into practical cognition so that it has some effect on what a cognitive agent does. There is a substantial literature in AI on multi-agent architectures that is potentially relevant to this. I suspect that a lot of light can be thrown on moral reasoning when viewed from this perspective.

5. Two Concepts of Rationality

The theory of rationality pertains to how an agent's cognition does or should proceed. In asking whether a person behaved rationally, we are asking whether he or she did did the right thing. The theory of rationality is the study of how to do it right when engaging in cognitive endeavors. The fundamental question is then, "What makes a procedure the right procedure for use in a particular cognitive context?"

5.1 Human Rationality

The question, "What makes rational procedures rational?" is made more difficult by the fact that there is more than one concept of rationality. We can illuminate one of these concepts by considering the way in which philosophers have traditionally tried to answer questions about how, rationally, to perform various cognitive tasks (e.g., reasoning inductively). The standard philosophical methodology has been to propose general principles, like the Nicod principle, or principles of Bayesian inference, or the hypothetico-deductive method, and test them by seeing how they apply to concrete examples. These are, in effect, thought experiments. We imagine reasoning in accordance with the principle in a certain context, and then consider whether that reasoning would be rationally correct.

Consider a concrete example. The *Nicod Principle* purports to describe inductive reasoning. It proposes that sets of premises of the form "(Ac & Bc)" confirm the generalization "All A's are B's", for any choice of A and B. There was a time when virtually all epistemologists endorsed the Nicod Principle in just this form. However, Goodman [1955] startled the philosophical world by constructing an example that was taken to conclusively refute this version of the Nicod Principle. This was his famous "grue" example. He defined:

> x is *grue* if and only if either (1) x is green and first examined before the year 2000, or (2) x is blue and not first examined before the year 2000.

Goodman then reasoned as follows. If we now (prior to the year 2000) examine lots of emeralds and find that they are all green, that gives us an inductive reason for thinking that all emeralds are green. Our sample of green emeralds is also a sample of grue emeralds, so if 'grue' were projectible then our observations would also give us a reason for thinking that all emeralds are grue. These two conclusions together would entail the absurd consequence that there will be no emeralds first examined after the year 2000. Clearly, we should not be able to draw this conclusion, so the Nicod Principle must be in error. Goodman proposed amending it by adding a qualification that A and B must be "projectible". He then proposed that "green" is projectible, but "grue" is not. This then led to a search for an analysis of projectibility, which continues to this day.

Goodman's refutation of the unrestricted Nicod Principle is conclusive, because everyone

who looks at the example agrees that it would not be rational to accept the conclusions drawn in accordance with the Nicod Principle. This illustrates the use of thought experiments in investigating rationality.

5.1.1 Rational Norms

Thought experiments provide data about what is rational or irrational in particular contexts, and then the philosopher constructs a general theory that is intended to capture that data. The theory consists of a set of *rational norms* compliance with which is supposed to constitute rationality. The thought experiments and the theory could be related in either of two ways. It is sometimes supposed that the thought experiments allow us to somehow "see" what the correct norms are, although it has never been explained exactly how that works. A more modest view would be that the norms are arrived at inductively on the basis of what our philosophical intuitions tell us about specific cases. On this view, philosophical theorizing about the contents of our rational norms would be indistinguishable from ordinary scientific theory confirmation. This inductive view of theorizing about rational norms would be the obvious choice if it were not for the fact that many philosophers endorse *cognitive essentialism*, according to which the correct norms governing various aspects of cognition are necessary features of the associated concepts, so the norms could not have been different. It follows from cognitive essentialism that the correct norms are necessarily correct. The theory of rationality is in the business of discovering necessary truths. It is generally supposed that necessary truths are *a priori*, so it would be at least odd to affirm cognitive essentialism and also claim that we discover the norms inductively. I will return to this issue below.

In order to test a principle by applying it to concrete examples, we must know how the example should come out. But how do people know that? The standard answer is "philosophical intuition", but that is not much of answer. What is philosophical intuition? If it is to support *a priori* conclusions about rational norms, it must be regarded as providing a window on a platonic universe of concepts and their relations. But this makes it very mysterious.

Considerable light can be thrown upon this methodology by comparing it with the methodology employed by linguists studying grammaticality. Linguists try to construct general theories of grammar that will suffice to pick out all and only grammatical sentences. They do this by proposing theories and testing them against particular examples. It is a fact that proficient speakers of a language are able to make grammaticality judgments, judging that some utterances are grammatical and others are not. These grammaticality judgments provide the data for testing theories of grammaticality. In describing what they are doing, linguists talk about appeals to "linguistic intuitions". The methodology seems to be parallel to the philosophical methodology, but in this case there is no temptation to suppose that the resulting theories of grammar are necessarily true or *a priori*. After all, they are describing a natural language. The structure of such a language is determined by convention and it changes over time. It is a contingent fact that a language has the grammatical rules it does, and if they were different our grammatical intuitions would have been different too.

5.1.2 Procedural Knowledge

When we learn a language, we learn *how to do* various things. Knowledge of how to do something is *procedural knowledge*. For most tasks, there is more than one way to do them. Accordingly, we can learn to do them in different ways, and so have different procedural knowledge. Procedural knowledge for how to perform a task can be described in terms of a set of norms for how to do it. When we acquire procedural knowledge, we acquire norms that, in some sense, have the power to control our behavior. Notice, however, that even when we know how to do something, we do not always comply with the procedural norms we have learned. For example, linguists observe that many, perhaps most, of our utterances are ungrammatical.

Our speech is populated with "Ahh"s and "Umm"s, we leave sentences unfinished, and commit a variety of other grammatical infractions. But we know better, so our procedural norms are not just a description of how we talk. In some sense, our procedural norms govern our behavior, but they do not do so in a way that guarantees that our behavior conforms to the norms.

How do procedural norms govern behavior? There is a model of this regulative process that is often implicit in philosophical thinking, but when we make the model explicit it is *obviously* wrong. This model assimilates the functioning of procedural norms to the functioning of explicitly articulated norms. For example, naval officers are supposed to "do it by the book", which means that whenever they are in doubt about what to do in a particular situation they are supposed to consult explicit regulations governing all aspects of their behavior and act accordingly. Explicitly articulated norms are also found in driving manuals, etiquette books, and so on. Without giving the matter much thought, there is a tendency to suppose that all norms work this way, and in particular to suppose that this is the way rational norms work. I will call this *the intellectualist model*.

It is noteworthy that epistemologists have often assumed something like the intellectualist model with regard to epistemic norms. It takes little reflection to realize that epistemic norms cannot function in accordance with the intellectualist model. If we had to make an explicit appeal to epistemic norms in order to acquire justified beliefs we would find ourselves in an infinite regress, because to apply explicitly formulated norms we must first acquire justified beliefs about how they apply to this particular case. For example, if we are to reason by making explicit appeal to a norm telling us that it is permissible to move from the belief that something looks red to us to the belief that it is red, we would first have to become justified in believing that the norm is included among our epistemic norms and we would have to become justified in believing that we believe that the object looks red to us. In order to become justified in holding those beliefs, we would have to apply other epistemic norms, and so on *ad infinitum*. Thus it is clear that epistemic norms cannot guide our reasoning in this way.

We have seen that is a purely logical reason why the intellectualist model cannot be a correct account of epistemic norms. Although it cannot be shown on logical grounds alone, the intellectualist model seems to be an equally bad model for the functioning of other procedural norms. Having procedural knowledge of what to do under various circumstances does not involve being able to give a general description of what we should do under those circumstances. This is the familiar observation that, for example, knowing how to ride a bicycle does not automatically enable one to write a treatise on bicycle riding. This is true for two different reasons. First, knowing how to ride a bicycle requires us to know what to do in each situation *as it arises*, but it does not require us to be able to say what we should do before the fact. Second, even when a situation has actually arisen, our knowing what to do in that situation need not be propositional knowledge. In the case of knowing that we should turn the handlebars to the right when the bicycle leans to the right, it is plausible to suppose that most bicycle riders do have propositional knowledge of this; but consider knowing how to hit a tennis ball with a tennis racket. I know how to do it—as the situation unfolds, at each instant I know what to do—but even at that instant I cannot give a description of what I should do. Knowing what to do is the same thing as knowing to do it, and that need not involve propositional knowledge.

We can give a rough description of how procedural norms govern behavior in a non-intellectualist manner. When we learn how to do something X, we "acquire" a plan of how to do it. That plan might (but need not) start out as explicit propositional knowledge of what to do under various circumstances, but then the plan becomes internalized. Using a computer metaphor, psychologists sometimes talk about procedural knowledge being "compiled-in". When we subsequently undertake to do X, our behavior is automatically channeled into that plan. This is just a fact of psychology. We form habits or conditioned reflexes. Norms for doing X constitute a description of this plan for doing X. The sense in which the norms guide our behavior in doing

X is that the norms describe the way in which, once we have learned how to do X, our behavior is automatically channeled in undertaking to do X. The norms are not, however, just descriptions of what we do. Rather, they are descriptions of what we *try* to do. Norms can be hard to follow and we follow them with varying degrees of success. Think, for example, of an expert golfer who knows how to swing a golf club. Nevertheless, he does not always get his stroke right. It is noteworthy, and it will be important later, that when he does not get his stroke right he is often able to tell that by something akin to introspection. When he does it wrong it "feels wrong". The ability to tell in this way whether one is doing something right is particularly important for those skills governing performances (like golf swings) that take place over more than just an instant of time, because it enables us to correct or fine tune our performance as we go along.

5.1.3 The Competence/Performance Distinction

The distinction between knowing how to do something and actually doing it generates what is called *the competence/performance distinction*. This originates with Chomsky [1957], who proposed that when the linguist constructs theories of grammar, she is trying to describe some of the procedural norms comprising speakers' procedural knowledge of how to speak the language. This is what is called *a competence theory* of language. It is contrasted with a *performance theory*, which is a description of what people actually do when they use language. A performance theory would include all the grammatical infractions as well as the grammatically correct utterances. Psychologists might be interested in performance theories of language, but the interest of the linguist is in competence theories.

It is noteworthy that procedural norms can be expressed using normative language. Consider knowing how to ride a bicycle. This procedural knowledge consists of knowing what to do under various circumstances, e.g., knowing to turn right when the bike leans to the right. The norm can equally be described as knowing what we *should* do under those circumstances. The point of using normative language to describe internalized norms is to contrast what the norms tell us to do with what we *do*. The simple fact of the matter is that even when we know how to do something (e.g., swing a golf club) we do not always succeed in following our norms. The point of the normative language is to highlight the competence/performance distinction. Note that an exactly similar use of normative language occurs in formulating rules of grammar. We are informed that under certain circumstances we should say 'that' rather than 'which'. The 'should' here is not a moral or prudential should. It is that kind of 'should' that we use in describing procedural knowledge.

5.1.4 Competence Theories of Cognition

My proposal is that the best way to understand the standard philosophical methodology for investigating rationality is to take it as parallel to the linguistic methodology for investigating grammaticality. We have (probably innate) procedural knowledge for how to perform various cognitive tasks. Our rational norms are the norms that describe this procedural knowledge. The way rational norms guide our cognition without our having to think about them is no longer mysterious. They describe an internalized pattern of behavior that we automatically follow in cognizing, in the same way we automatically follow a pattern in bicycle riding. This is what rational norms are. They are the internalized norms that govern our cognition. Once we realize that they are just one more manifestation of the general phenomenon of automatic behavior governed by internalized norms, rational norms should no longer seem puzzling. The mystery surrounding rational norms evaporates once we recognize that the governing process is a general one and its application to rational norms and reasoning is not much different from its application to any other kind of procedural norms. Of course, unlike most norms our rational norms are probably innate, in which case there is no process of internalization that is required to make them available for use in guiding our reasoning.

The strongest argument for this understanding of rational norms is that it makes our standard philosophical methodology understandable. It was observed that theories of rationality are defended by appeal to philosophical intuitions. This is similar to the way linguists use linguistic intuitions to defend theories of grammar. What are these philosophical and linguistic intutions? A general characteristic of procedural knowledge seems to be that once we have it, we can also judge more or less reliably whether we are conforming to it in particular cases. For example, once I have learned how to ride a bicycle, if I lean too far to one side (and thus put myself in danger of falling), I do not have to wait until I fall down to know that I am doing it wrong. I can detect my divergence from what I have learned and attempt to correct it before I fall. Similarly, it is very common for competent speakers of a language to make ungrammatical utterances. But if they reflect upon those utterances, they have the ability to recognize them as ungrammatical and correct them. In other words, at least in human beings, procedural knowledge carries with it a general ability to monitor performance and detect divergences from learned norms. This is a kind of introspection, and its usefulness is clear. For example, in swinging a tennis racket I monitor my swing, I can tell at each instant whether I am too high or too low, and I can correct my swing accordingly. Thus it is important to the operation of procedural norms that compliance with them be introspectible (although not necessarily with total reliability).

This ability to detect divergences from learned procedural norms is arguably what the linguist is employing in making judgments of grammaticality. She imagines speaking in a certain way, and then considers whether that would conform with her grammatical norms. If not, she marks it as ungrammatical, and uses that as a datum for her theory of grammaticality. Linguistic intutions are the output of this divergence detection mechanism as applied to language production.

I propose that philosophical intuitions are to be understood in precisely the same way. We know how to cognize, and by virtue of having that procedural knowledge we can also detect (with varying reliability) divergences from the procedural knowledge underlying rational cognition. The intuitions to which philosophers appeal are the result of this divergence detection mechanism. The argument for this view is that it takes the mystery out of our philosophical intuitions, and makes it completely explicable what we are up to in constructing theories of rationality. Thus the theories of rationality that philosophers construct by appealing to their intuitions about rationality are best viewed as competence theories of cognition. That is, they are attempts to articulate the rules comprising our procedural knowledge for how to cognize. This makes them completely parallel to linguistic theories of grammar.

Unlike linguistic knowledge, it seems pretty clear that large parts of our procedural knowledge of how to cognize are built into us rather than learned. This may be required as a matter of logic—it may prove impossible to get started in learning how to cognize unless we already know how to cognize to some extent. But even if that is not true, it is overwhelmingly likely that evolution has built into us knowledge of how to cognize so that we do not come into the world so epistemically vulnerable as we would otherwise be. This way, we at least know how to get started in learning abut the world. This is rather strongly confirmed by the overwhelming agreement untutored individuals exhibit in their procedural knowledge of how to cognize.[5] For example, psychological evidence indicates that everyone finds reasoning with modus ponens to be natural and reasoning with modus tollens to be initially unnatural.[6] It is unlikely that this is something we have learned.

---

[5] This is not to say that everyone is equally good at cognizing. Cognitive performance varies dramatically. But insofar as people make cognitive mistakes, they can generally be brought to recognize them as such, suggesting that their underlying procedural knowledge is the same.

[6] See Wason [1966] and Cheng and Holyoak [1985].

Notice that the rules comprising our procedural knowledge of how to cognize cannot be viewed as mere generalizations about how we do cognize. That would make them descriptions of cognitive performance rather than cognitive competence. Instead, they are the rules that we, in some sense, "try" to comply with. They are the rules perceived divergence from which leads us to correct our cognitive performance to bring it into compliance.

Although competence theories are importantly different from performance theories, they are still about something that is "psychologically real" and a contingent feature of the human beings that possess the procedural knowledge being described. In particular, it is a contingent fact that the human cognitive architecture is build the way it is, even if its structure is innate rather than learned. So in constructing competence theories of human cognition, we are describing contingent features of human beings. This is, in principle, the sort of thing that psychologists investigate. However, the current methodology of cognitive psychology is better suited for constructing performance theories than competence theories. It is difficult to see how to get beyond what people do and instead study what they know how to do. As cognitive psychology matures, this may change, but in the meantime the best methodology for studying human cognitive competence is probably the traditional methodology of the philosopher. In precisely the same way, the study of English grammar is about psychologically real features of English speakers, but the best way to study it is probably the traditional method of the linguists which appeals to linguistic intuitions.

One of the initial puzzles about rationality is why some higher cognitive processes are regarded as falling under the purview of rationality, but not others. For example, visual processing is excluded but reasoning is included. This account of rationality as "cognitive competence" suggests a solution to this puzzle. The distinction between rational and non-rational cognitive processes seems to have to do with the possibility of introspective monitoring. The rational assessment of a cognitive performance requires monitoring the performance and evaluating whether it conforms with our procedural norms. If we cannot monitor it, we cannot evaluate it, and then there seems no point to saying that it is governed by procedural norms rather than just saying that there are general principles that describe how the process works. Visual processing is a black box—we cannot introspect what goes on there. The course of reasoning, on the other hand, is introspectible, so it falls within the purview of rationality.

5.1.5 Using Philosophical Intuitions

Philosophers have often supposed that our philosophical intuitions provide us with *a priori* insights into the functioning of particular rational norms. We can now see why they thought that. Our philosophical intuitions are the result of a kind of introspective access to whether we are complying with our procedural norms. As such, they are quite different from the kind of access an empirical psychologist might have by performing some kind of laboratory experiment on large numbers of undrgraduates. But it is probably misleading to call these intuitions "*a priori*". They are more like the proprioceptive access we have to the positions of our limbs. That is quite different from judging the positions of our limbs by looking in a mirror, but it isn't *a priori*.

How do we use the compliance data we acquire from our philosophical intuitions in constructing general theories of the contents of our rational norms? It was once a common view in analytic philosophy that our philosophical intuitions allow us to just "see" what the correct norms are. This might be somewhat plausible for simple epistemic norms like modus ponens. But it is hard to see how our philosophical intuitions could reveal the norms themselves rather than just instances of them, and it seems wrong anyway when we consider the case of complicated norms. Consider, for example, the norms governing inductive reasoning. Once, perhaps, people thought they were simple, being correctly described by something no more complex than the Nicod principle. However, more recent work on induction reveals that the correct form of the

norms governing inductive reasoning is both elusive and complex, and at this point no one can say with confidence that they know exactly what they are. Philosophical intuitions provide us with copious examples of correct and incorrect inductive reasoning, but formulating norms that capture those intuitions proves extremely difficult. That task seems completely analogous to constructing a scientific theory to explain a large set of data. I think it should be concluded that our philosophical methodology here is just general scientific theory confirmation. Our philosophical intuitions provide us with data, and as philosophers we try to construct procedural norms that would explain that data. The process is broadly inductive.

5.1.6 Cognitive Essentialism

It was remarked above that it is a contingent fact what rational norms are built into the human cognitive architecture. We might have been built differently than we are. So does it follow that those rational norms are contingent, and cognitive essentialism is wrong? This is controversial, but the answer need not be that cognitive essentialism is false. The key to defending cognitive essentialism is to endorse a conceptual role theory of mental content and concept individuation. On such a theory, what concepts are expressed by a cognitive agent's system of mental representations is determined, at least in part, by the rational norms governing the agent's cognition. Thus, for example, part of what makes a mental representation represent redness is that the agent's epistemic norms license a defeasible inference from an object's looking red to its being in the state represented. A theory of this sort was defended in my [1989 (*How to Build a Person*)]. Rather than try to defend it again here, I will just suggest it as a possible theory of mental content. This view has the consequence that the rational norms associated with a concept are a necessary feature of the concept, because they are part of what make it the concept that it is. So cognitive essentialism is true on this view, even though it is a contingent fact that the human cognitive architecture is based on the rational norms it is. The contingency translates into a contingency regarding which concepts human beings employ, rather than a contingency regarding what norms attach to what concepts.

Although, on the view being considered, it is a necessary truth that the concepts we employ are governed by the rational norms we employ, this is a *de re* necessity, not a *de dicto* necessity. That is, the concepts are necessarily such that they are governed by those rational norms. As it is only a *de re* necessity, there is no reason to expect a statement of it to be true *a priori*. Statements of *de re* necessity are typically *a posteriori* truths.

5.1.7 Human Irrationality

It is a consequence of this account of rationality that humans are in principle capable of behaving rationality, because rational norms encode our procedural knowledge of how to cognize. If for either psychological or computational reasons, human cognizers could not comply with a proposed norm, it follows that that norm is not part of their procedural knowledge for how to cognize, and hence that it is not one of the procedural norms comprising human rationality.

A surprising feature of a number of philosophical theories of rationality is that they would make it impossible for humans, or for that matter any other resource bounded (i.e., "real") cognitive agents to behave rationally. For example, a not uncommon view is that a correct theory of practical rationality should be based on decision theory. A common construal of decision theory has it that in choosing what actions to perform, a rational agent should consider all possible actions and select one that maximizes expected utility. But this is clearly impossible. There are always infinitely many possible actions. No resource bounded agent can consider them all, compute their expected utilities, and then select an optimal one. Such a rational norm could not be complied with. But then it follows that it is not one of the norms comprising human rationality. Some similar more heavily qualified norm probably is, but this particular norm has been stated too simply to be a viable candidate for a principle of human

rationality.

Perhaps most philosophical theories of various aspects of rationality have the consequence that it is psychologically or computationally impossible for real cognitive agents to comply with them. For example, you often hear philosophers asserting that a rational agent should believe all the logical consequences of its beliefs. But it is generally recognized that it is computationally impossible to build an agent that could comply with that norm, so that cannot be a correct rational norm. A correct rational norms that appeals to logical consequences must be heavily qualified. It might, for example, say something to the effect that a rational agent should update its beliefs in light of newly discovered logical consequences.

This is a general difficulty for philosophical theories of rationality, and it means that most such theories cannot be taken seriously as contenders for the true account of rationality. However, this is not to say that they are without value. It will be suggested below in connection with both epistemic and practical cognition, that traditional theories of rationality are often better viewed as theories of of what epistemic or practical conclusions a cognitive agent should *end up with* if it could perform all possible relevant cognizing and then step back and consider the result. In a sense to be made precise below, such theories are theories of epistemic or practical *warrant* rather than epistemic or practical *justification*. Such theories do not provide viable accounts of rational cognition, but they can be useful nevertheless in describing a kind of "ideal target" at which rational cognition aims.

It follows from this account of rationality that humans are fundamentally rational, at least in the sense that they know how to behave rationally even if they sometimes fail to do it. This might seem dubious in light of the fact that there has been a lot of recent work in psychology aimed at establishing that humans exhibit pervasive patterns of irrationality.[7] That work is not all equally successful. For example, much of it is concerned with showing that people tend to make mistakes in complex probability judgments, but that does not establish irrationality any more than showing that people tend to make mistakes in the integral calculus would. However, other examples are more persuasive. For example, Wason [19??] describes an experiment in which subjects are presented with four cards, two face up and two face down. The visible sides display red, black, spade, and diamond. The subjects are asked which cards they must turn over to test the hypothesis that all the red cards are diamonds. The vast majority of subjects respond that they must turn over the red card and the diamond. However, turning over the diamond would not show anything, because even if it were black, that would not disprove the hypothesis. On the other hand, the spade must be turned over, because if it were red then the hypothesis would be false. So the right answer is that one should turn over the red card and the diamond. People get it wrong, and it is tempting to say that because this is such a simple case it demonstrates irrationality. I am not entirely convinced that this is a case of irrationality, but even if it is, it does not show that people do not *know how* to solve this problem. It shows that they do not in fact solve it correctly, but that is a matter of human performance rather than human competence. That people do have procedural knolwedge adequate for solving the problem follows from the fact that they recognize that they got it wrong when the proceeding considerations are called to their attention, and they see what the correct answer is. Although we sometimes have difficulty reasoning correctly, we *can* do it. If our procedural knowledge did not enable us to do it right, we would not be able to recognize that people have a tendency to do it wrong. All this really illustrates is the how hard it is to study human competence, rather than human performance, using the standard tools of cognitive psychology.

---

[7] Much of the psychological material can be found in Daniel Kahneman, Paul Slovic, and Amos Tversky [1982], and R. E. Nisbett and L. Ross [1980].

5.2 Generic Rationality

I have proposed that the best way to understand the standard philosophical methodology for investigating rationality is to take it as an attempt to elicit the procedural norms comprising a competence theory of human cognition. I will refer to this concept of rationality as *human rationality*. Theories of human rationality are specifically about human cognition. They describe contingent psychological features of the human cognitive architecture. However, upon reflection this seems to be a rather parochial view of rationality. The human cognitive architecture is a product of evolution. Environmental pressures have led to our evolving in particular ways. We represent one solution to various engineering problems that were solved by evolution. But there is no reason to think that these problems always have a single, or even a single best, solution. Some features of human rationality may be strongly motivated by general constraints imposed by the design problem rationality is intended to solve, but other features may be quite idiosyncratic, reflecting rather arbitrary design choices. To illustrate this with a simple example, it was remarked above there is overwhelming psychological evidence that human beings do not employ *modus tollens* as a primitive inference rule. They can learn to use it, but it is not built into their system of cognition from the start. If we could construct an artificial agent that was the cognitive duplicate of human beings except that it also employed *modus tollens* as a primitive inference rule, we would not regard its cognitive behavior as irrational, despite the fact that it would not conform exactly to human norms of rational cognition. Standard philosophical methodology leads to an anthropocentric view of rationality. An agent whose procedural knowledge for how to cognize differed from that of human beings would not automatically be irrational. This seems to indicate the need for a more general "generic" concept of rationality. How can we understand this concept?

We can think of rationality as the solution to certain design problems. I will argue below that quite general features of these problems suffice to generate much of the structure of rationality. General logical and feasibility constraints have the consequence that there is often only one obvious way of solving certain design problems, and this is the course taken by human rationality. Much of the general structure of human rationality can be explained in this way. However, the *details* of the design of human cognition are not similarly forthcoming. It may often be the case that the details could be filled out in several different ways, and human cognition represents just one choice for how to do that. There could be other rational agents, either natural or artificial, whose cognitive architectures represent somewhat different solutions to the same design problems.

Generic rationality emerges from approaching rational cognition from the design stance. Rationality represents the solution to certain design problems. Approaching rationality from the design stance is more common in artificial intelligence than it is in philosophy. In philosophy, the emphasis has been on human rationality, but in artificial intelligence there has been less concern with mimicing human rationality and more interest in designing intelligent (rational) systems that will accomplish certain cognitive tasks in humanlike ways but not necessarily in exactly the way humans do it. The details of human rational cognition may be of particular interest to cognitive psychologists, but it is hard to see why they should be of great interest to philosophers. It seems to me that philosophy should be interested in rational cognition in general. How is rational cognition possible, and how might it work? The two may not be all that different if many of the principles of human cognition represent the only solution to elements of the design problem underlying generic rationality. However, where the design of human rationality represents arbitrary design decisions, they will differ. In that case it strikes me as often more interesting to study arbitrary rational agents. An anthropocentric bias makes sense if you are doing psychology, but why should philosophy be so constrained? Thus the focus of this investigation will be on generic rationality rather than specifically human rationality. It is to be expected, however, that studying rationality in general will throw considerable light on why the human cognitive architecture has the design it does. Conversely, the best way to solve a problem

in cognitive engineering may be to see how it is solved in an existing system, and for most sophisticated cognitive tasks, the only existing system we can look at is human beings. So even if our interest is generic, we will typically find ourselves drawing insight from an inspection of specifically human rational cognition.

5.2.1 The Doxastic-Conative Loop

Approaching rationality from the design stance, I propose to understand generic rationality as attaching to any agent that constitutes a solution to the design problem that also generates the human cognitive architecture. However, if we are to explain rationality in this way, we must say what design problem rationality is designed to solve. I will begin by exploring one possibility, and then generalize the account.

A simplistic view of evolution has it that evolutionary pressures select for traits that enhance the survivability of the creature. So we might regard the design problem for rationality to be that of creating an agent that can survive in a hostile world *by virtue of its cognitive capabilities*. This presupposes a prior understanding of what cognition is and how it might contribute to survivability. I take it as characteristic of rational cognition that a cognitive agent has doxastic states ("beliefs", broadly construed) reflecting the state of its environment, and conative states evaluating the environment as represented by the agent's beliefs. It is also equipped with cognitive mechanisms whereby it uses it beliefs and conations to select actions aimed at making the world more to its liking. This comprises the *doxastic-conative loop*, as diagrammed in figure 1.1.

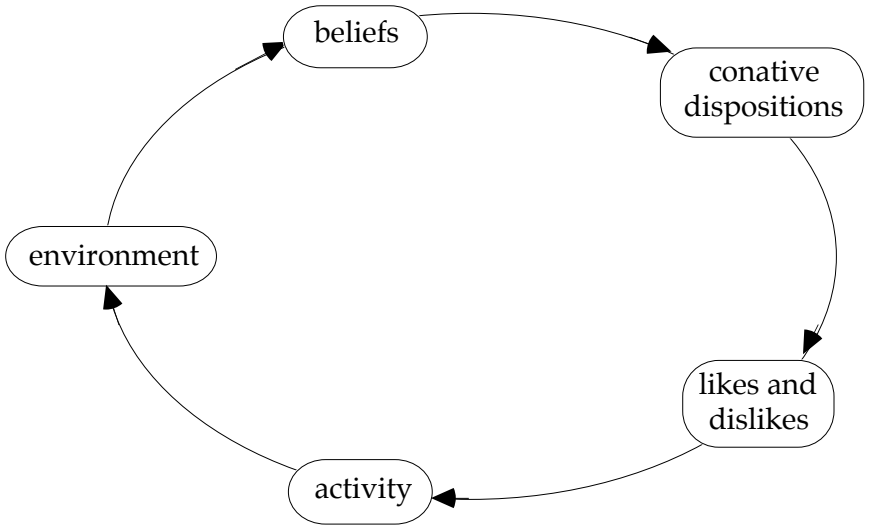

Figure 1.1 Doxastic-conative loop

A rational agent has beliefs reflecting the state of its environment, and it likes or dislikes its situation. When it finds the world not entirely to its liking, it tries to change that. Its cognitive architecture is the mechanism whereby it chooses courses of action aimed at making the world more to its liking. Within the doxastic-conative loop, epistemic cognition is, in an important sense, subservient to practical cognition. The principal function of cognition is to direct activity (practical cognition), and the role of epistemic cognition in rationality is to provide the factual background required for practical cognition.

Let us define a *cognitive agent* to be one implementing the doxastic-conative loop. I take it that rationality pertains to cognitive agents. We can assess a cognitive architecture in terms of how well it achieves its design goal. However, different design goals are possible. As remarked above, a natural design goal is to construct an agent capable of using its cognitive capabilties to survive in an uncooperative environment. This is motivated by considerations of evolution and natural selection, but it is based upon a simplistic view of evolution. For biological agents, a more natural design goal is propagation of the genome. And for artificial agents, we may have many different design goals. For example, in designing a Kamakaze robot warrior, the objective is neither survival nor propogation of the genome. This suggests that there is no privileged design goal in terms of which to evaluate cognitive architectures.

It is striking, however, that for agents that implement the doxastic-conative loop, the same cognitive features seem to contribute to achievement of a wide range of design goals. For example, for many different choices of design goal, an agent operating in a complex and uncooperative environment and intended to achieve its design goal by implementing the doxastic-conative loop will probably work better if it has a rather sophisticated system of mental representation, is capable of both deductive and defeasible reasoning, and can engage in long-range planning. This is equally true of biological agents and Kamakaze robot warriors. This suggests that there is a relatively neutral perspective from which we can evaluate cognitive architectures, and this in turn generates a generic concept of rationality.

5.2.2 Truth-Conduciveness

The evaluation of cognitive architectures in terms of how well they achieve their design goals is reminiscent of externalist epistemology. One prominent version of reliabilism (called "norm reliabilism" in Pollock [1987]) evaluates epistemic norms in terms of their truth-conduciveness. I argued that norm reliabilism fails as applied to human epistemic norms, because our norms already accommodate without revision the kinds of changes to reasoning that norm reliabilism dictates. But perhaps where truth-conduciveness really comes in is in the evaluation of general cognitive architectures. If a rational agent is to use its beliefs to guide it choice of actions, it does seem superficially desirable that it have true beliefs. Let us consider this more carefully.

The first thing to notice is that if truth-conduciveness is to be a desirable property of cognitive agents, this must be derivative from its contribution to the satisfaction of the agent's design goals. Those design goals are unlikely to explicitly include the possession of true beliefs. Rather, cognitive agents aim at achieving various practical goals, and true beliefs will be valuable only insofar as they contribute to the achievement of those practical goals. The entire cognitive architecture must be evaluated as a package. The evaluation of the specifically epistemic parts of the architecture must be derivative from their contribution to the practical success of the entire architecture.

How, exactly, should truth enter into the evaluation of a cognitive architecture? There are several possibilities. Perhaps the desideratum should be that cognitive agents have lots of true beliefs. However, this cannot be right because it could be achieved by simply believing everything. That would not contribute to practical success. Perhaps, instead, the desideratum should be that the agent have *only* true beliefs. That, however, could be achieved by believing nothing. Again, that will not contribute to practical success. An initially more plausible idea is that we want the agent to simultaneously maximize true belief and minimize false belief, or better, we want the agent to maximize the ratio of true beliefs to false beliefs. But this could be achieved by believing every tautology of the form $(P \vee \sim P)$ and nothing else. Again, this will not contribute to practical success.

The last example illustrates that it is not truth per se, but *interesting* truth that the agent should be pursuing. Interesting truths are those that are of particular use in practical cognition,

e.g., in deciding how to act so as to achieve one's goals. But even here, it is not clear to what extent truth is the relevant desideratum. The agent should pursue beliefs that will help it be effective in achieving its goals, and in some cases it seems clear that truth is required, For example, it seems desirable to have true beliefs about whether a tiger is about to eat you. But in other cases it is not so clear that true beliefs will be the most helpful. True beliefs might be too complex to be used efficiently in day to day deliberation. Approximately true beliefs that are simple may be more useful. Consider deciding what route to take while walking across your front lawn to your door. If you had to do that by solving a problem in quantum mechanics, you would have a terrible time getting into your house. It is much more useful to have a number of approximately true "rules of thumb" that you can use in making such decisions, e.g., "If you try to walk over the tricycle rather than around it you are apt to fall down."

The upshot is that it is not clear how the pursuit of true belief should enter into the evaluation of cognitive architectures. It seems clear that truth is often a good thing, but not all truths are equally desirable, nor are all falsehoods equally undesirable. The relationship of truth to rationality is not a simple matter.

6. Artificial Intelligence, Philosophy, and Rationality

Philosophy and psychology are not the only disciplines with traditional interest in rational cognition. The investigation of rationality is also of importance for artificial intelligence. However, its importance varies for different parts of AI. AI is not a monolithic discipline. Work in AI is directed at several different goals:

- Sometimes the objective of AI researchers is to model human cognition. That is basically a psychological endeavor, and philosophical theories of rationality do not seem to be directly connected to it.
- Sometimes the objective is to model human *rational* cognition. This is basically a philosophical endeavor, with psychological overtones.
- Perhaps the original objective of AI was to build an autonomous rational agent—an intelligent robot with humanlike capabilities. It was not assumed that the cognition of such an agent would exactly model human rational cognition, but it would nevertheless be rational in the generic sense.
- Most often, AI researchers have been concerned to solve practical problems of automated data processing and decision making. If these problems are sufficiently narrowly constrained, they can be addressed as engineering problems. This assumes, however, that we know what answers we want our AI system to produce in particular cases, so that we know what we are trying to get the system to do and can tell whether it is doing it correctly.
- As the problems of automated data processing and decision making become more general and the system is expected to operate in less constrained environments, it becomes increasingly difficult to know just what we want the system to do. The specification of the engineering problem becomes more and more difficult. We want the system to make the *right* decisions and draw the *right* conclusions, but to design a system that will do this in a wide variety of contexts we need a general description of what constitutes the right decisions and the right conclusions. Basically, we want the conclusions and decisions to be *rational*. Thus as the range of applicability of the AI system is broadened, the practical AI problem gets closer and closer to the problem of either modeling human rationality or building an autonomous rational agent. So we need a general account of rationality.

6.1 AI Needs Philosophy

Apparently many of the more sophisticated AI endeavors require a theory of rationality. If the objective is to model human rationality, this is obvious. But it is no less true of the more general project of constructing autonomous rational agents, or many of the more applied projects of automatic data processing and decision making in complex environments. The latter do not exactly require a theory of human rationality, but they do require a theory of generic rationality and the construction of such a theory is a philosophical task. Furthermore, in constructing autonomous rational agents, although the design goal need not be to build a system that models human rational cognition, the problem of constructing artificial systems that perform complex cognitive tasks is sufficiently difficult that the best way to construct such a system will often be to figure out how human cognition performs the tasks and then model the artificial system at least roughly on that working model. Furthermore, if, as I have suggested, logical and computational constraints often have the consequence that sophisticated features of human cognition, like the ability to reason defeasibly, represent the only or the only obvious way to solve certain cognitive problems, it follows that these cognitive features must be shared by all sophisticated autonomous rational agents.

6.2 Philosophy Needs AI

In the previous section, I argued that sophisticated AI must be based upon philosophical theories of rationality. In other words, AI needs philosophy. I believe it is equally true that philosophy needs AI. Philosophical theories of anything tend to be armchair theories. Philosophers propound them and test them by just sitting and thinking about them. Sometimes the philosopher's reflections are informed by empirical information from other disciplines, but the philosopher's contribution generally consists of just sitting and thinking. This reflects a certain conception of what philosophy is all about. Philosophy is "non-empirical", in the sense that the philosopher does not go out and collect empirical data in the course of his or her philosophizing. I have no argument with this conception of philosophy, but I do have an argument with the methodology it has spawned. It is my conviction that armchair epistemology is not, in general, very reliable. The difficulty is twofold.

First, there is little incentive to get the details precisely right, because nobody will notice anyway. Impressionism often seems more appealing than photorealism. That would not be so bad if we could be confident that we *can* always get the details right if we just put a bit more work into it. But in fact, when philosophical theories fail it is usually because the details cannot be made to work. Grand pictures painted with broad brushstrokes are fine for hanging on the wall and admiring for aesthetic reasons, but if the objective is to discover truth, it is essential to see whether the details can be made to work. One way to make sure that the details have been provided is to try to build an AI system that implements the theory. This requires the theory to be sufficiently precise that it can actually be implemented, and it is amazing how effective that is at uncoverying gaps in the theory that previously went unnoticed. It is remarkably common when implementing a theory to discover to your chagrin that there are significant parts of the theory that you simply overlooked and forgot to construct. To the armchair bound philosopher, that may sound remarkably stupid, but that is only because he has never tried implementing his theories and has thereby never had the opportunity to make the same humbling discoveries about his own thought.

This brings us to the second problem. Theories constructed and tested in the armchair can only be tested by seeing how they apply to very simple examples, because these are the only examples we are able to work through just by thinking about them. The armchair philosopher, in his naivete, may suppose that if a theory is going to fail it will fail on simple examples. That is just not true. AI has a long history (long for AI at least) of constructing theories and testing them

on "toy problems". For example, much early AI work on planning was tested on the blocks world, which is a world consisting of a table top with children's blocks scattered about and piled on top of each other, and the planning problems were problems of achieving certain configurations of blocks. AI learned the hard way that systems that worked well for such toy problems frequently failed to scale up to problems of realistic complexity. There is every reason to expect the same thing to be true of philosophical theories of rational cognition. The only way to give them a fair test is to implement them and apply them to problems of real-world complexity.

It is worth mentioning a third problem. My personal experience has been that if our only computational tool is our armchair, we will often fail to solve even the toy problems despite our thinking that we have done so. This is because it can be very difficult to work out all of the consequences of a complex philosophical theory even as applied to a toy problem.

The solution is for the philosopher to join forces with AI. If a theory of rational cognition is formulated with sufficient precision that it can be implemented, then we can make use of the computational power of the computer to apply it to concrete examples and see whether it does what we expect. My experience has been that implemented theories almost always fail to work as expected the first time around. Implementation is a fast way to find gaps in theories, hidden incoherences, and counterexamples. The course of my own philosophical investigations has been profoundly affected by implementing my theories. In many cases I have been led to change my theories radically or reject them altogether.

A theory of rational cognition is a theory of how to achieve certain cognitive tasks. One part of such a theory is the claim that a certain abstractly described functional architecture will work in the manner envisioned and solve the desired cognitive problems. If the theory is a theory of human rationality, it has a second part which claims that human cognition works in the manner described. Implementing a theory of rational cognition can prove that it is correct as a theory of generic rationality. The implementation provides the vehicle for establishing that the system does achieve the cognitive goals in the manner envisioned. Of course, this does not also establish that human cognition works in that way. However, the implementation can provide indirect evidence for the claim that human rational cognition works in the manner described. Human rational norms describe our procedural knowledge for how to cognize. As such, a necessary condition for an account of rational norms to correctly describe human cognitive competence is that it be possible for a cognitive agent to actually work that way. That turns out to be a difficult condition to satisfy. It is hard to construct *any* system capable of performing sophisticated cognitive tasks. Consequently, if there is independent reason to think that human cognition works in a certain rather general way, constructing a precise system conforming to the general description and showing that the system actually works provides some evidence for the claim that human cognition works that way, or at least in a very similar way.

6.3 The OSCAR Project

The upshot is that sophisticated AI systems, with the exception of those aimed at the psychological modeling of human cognition, must be based on general philosophical theories of rationality, and conversely, philosophical theories of rationality should be tested by implementing them in AI systems. So the philosophy and the AI go hand in hand. These considerations motivated the beginning of the OSCAR Project in the mid 80's. The objective of the OSCAR Project is twofold. On the one hand it is to formulate a general theory of rationality, both epistemic and practical. On the other hand, it is to implement the theory thus formulated, and use the implementation to apply the theory to concrete examples. The result has been the OSCAR architecture for rational agents, described in my [1995] and further expanded and refined in this book. The core of the architecture is an implemented system of defeasible reasoning that has its origins in my epistemological work dating back to the late 60's. This system of defeasible reasoning provides the inference engine for an autonomous agent. Philosophical theories of

specific kinds of defeasible reasoning, e.g., reasoning from perceptual input and certain kinds of temporal and causal reasoning, generate reason-schemas whose implementation extends the capabilities of the agent. For example, in this book I will describe an implementation of a solution to the frame problem. Similarly, philosophical work on practical reasoning described later in the book has led to an implemented theory of defeasible planning that extends, in various ways, conventional AI planning technology.[8] Work is currently underway on refining and implementing the theory of probabilistic and inductive reasoning propounded in my [1990]. Of course, there are more problems remaining to be solved than there are problems for which I have even tentative solutions to propose. But that is the nature of the scientific enterprise. I suspect that, ultimately, virtually all familiar philosophical problems will turn out to be at least indirectly relevant to the task of building an autonomous rational agent, and conversely, the AI enterprise has the potential to throw light at least indirectly on most philosophical problems.

---

[8] Pollock [1998a].