2/28/06

Irrationality and Cognition

John L. Pollock
Department of Philosophy
University of Arizona
Tucson, Arizona 85721
pollock@arizona.edu
http://www.u.arizona.edu/~pollock

Abstract

The strategy of this paper is to throw light on rational cognition and epistemic justification by examining irrationality. Epistemic irrationality is possible because we are reflexive cognizers, able to reason about and redirect some aspects of our own cognition. One consequence of this is that one cannot give a theory of epistemic rationality or epistemic justification without simultaneously giving a theory of practical rationality. A further consequence is that practical irrationality can affect our epistemic cognition. I argue that practical irrationality derives from a general difficulty we have in overriding built-in shortcut modules aimed at making cognition more efficient, and all epistemic irrationality can be traced to this same source.

A consequence of this account is that a theory of rationality is a descriptive theory, describing contingent features of a cognitive architecture, and it forms the core of a general theory of "voluntary" cognition — those aspects of cognition that are under voluntary control. It also follows that most of the so-called "rules for rationality" that philosophers have proposed are really just rules describing default (non-reflexive) cognition. It can be perfectly rational for a reflexive cognizer to break these rules.

The "normativity" of rationality is a reflection of a built-in feature of reflexive cognition — when we detect violations of rationality, we have a tendency to desire to correct them. This is just another part of the descriptive theory of rationality.

Although theories of rationality are descriptive, the structure of reflexive cognition gives philosophers, as human cognizers, privileged access to certain aspects of rational cognition. Philosophical theories of rationality are really scientific theories, based on inference to the best explanation, that take contingent introspective data as the evidence to be explained.

1. The Puzzle of Irrationality

Philosophers ask, "What should I believe? What should I do? And how should I go about deciding these matters?" These are questions about rationality. We want to know how we, as real cognizers, with all our built-in cognitive limitations, should go about deciding what to believe and what to do. This last point deserves emphasis. Much work on rationality is about "ideal rationality" and "ideal agents". But it is not clear what ideal rationality has to do with real, resource-bounded, agents. We come to the study of rationality wanting to know what *we* should do, and this paper is about that concept of rationality.

Philosophers, particularly epistemologists, regard irrationality as the nemesis of the cognizer, and they often think of their task as that of formulating rules for rationality. Rules for rationality are rules governing how cognitive agents should perform their cognitive tasks. If asked for simple examples, philosophers might propose rules like "Don't hold beliefs for which you do not have good reasons (or good arguments)", "When you do have a good argument for a conclusion, you should accept the conclusion", and "Be diligent in the pursuit of evidence". Epistemological theories are often regarded as proposing more detailed rules of rationality governing things like inductive reasoning, temporal reasoning, and so forth, and theories of practical reasoning propose rules for rational decision making.

Philosophers seek rules for avoiding irrationality, but they rarely stop to ask a more fundamental question. Why is it possible for humans to be irrational? We have evolved to have a particular cognitive architecture. Evolution has found it useful for us to reason both about what to believe and about what to do. Rationality consists of reasoning, or more generally, cognizing, *correctly*. But if rationality is desirable, why is irrationality possible? If we have built-in rules for how to cognize, why aren't we built to always cognize rationally? Consider the steering mechanism of a car. There are "rules" we want it to follow, but we do that by simply making it work that way. Why isn't cognition similar? An even better comparison is with artificial cognitive agents in AI. For example, my own system OSCAR (Pollock 1995) is built to cognize in certain ways, in accordance with a theory of how rational cognition ought to work, and OSCAR cannot but follow the prescribed rules. Again, why aren't humans like this? Why are we able to be irrational?

The simple answer might be that evolution just did not design us very well. The suggestion would be that we work in accordance with the wrong rules. But this creates a different puzzle. If we are built to work in accordance with rules that conflict with the rules for rationality, how does rationality come to have any psychological authority over us? In fact, when we violate the rules of rationality, and subsequently realize we have done so, we feel a certain pressure to "correct" our behavior and conform to the rules. However, if we are built to act in accordance with rules that lead to our violating the rules of rationality, where does this pressure for conforming to them come from? Their authority over us is not just a theoretical authority described by philosophical ideals. They have real psychological authority over us. From whence do they derive their authority? If evolution could build us so that rationality has this kind of *post hoc* authority over us, why could it not have built us so that we simply followed the rules of rationality in the first place?

It cannot be denied that we are built in such a way that considerations of rationality have psychological authority over us. But we are not built in such a way that they have absolute authority — we can violate them. What is going on? What is the role of rationality in our cognitive architecture? Why would anyone build a cognitive agent in this way? And by extension, why would evolution build us in this way?

These puzzles suggest that we are thinking of rationality in the wrong way. I am going to suggest that these puzzles have a threefold solution. First, the rules philosophers have typically proposed are misdescribed as “rules for rationality”. They play an important role in rational cognition, but it can be perfectly rational to violate them. Second, the reason we can violate them is that we are reflexive cognizers who can think about our own cognition and redirect various aspects of it, and there are rules for rationality governing how this is to be done. But, third, we do still behave irrationally sometimes, and that ultimately is to be traced to a particular localized flaw in our cognitive design.

Having proposed an account of irrationality, we will be able to use that to throw light on rationality. I will urge that the task of describing the rules for rationality is a purely descriptive enterprise, of the sort that falls in principle under the purview of psychology. Still, there is something normative about the rules for rationality, and I will try to explain that. Although, on this account, theories of rationality are empirical theories about human cognition, the nature of reflexive cognition provides philosophers with a privileged access to rational cognition, enabling us to investigate these matters without performing laboratory experiments.

First, some preliminaries.

2. Rationality, Epistemology, and Practical Cognition

Much of epistemology is about how beliefs should be formed and maintained. It is about “rational doxastic dynamics”. Beliefs that are formed or maintained in the right way are said to be *justified*. This is the “procedural” notion of epistemic justification that I have written about at length in my works in epistemology (Pollock 1987, 1997; Pollock and Cruz 1999). It is to be contrasted with the notions of epistemic justification that are constructed for the sake of analyzing “*S* knows that *P*”, an enterprise that is orthogonal to my present purposes.

Procedural epistemic justification is closely connected to rationality. We can distinguish, at least loosely, between epistemic cognition, which is cognition about what to believe, and practical cognition, which is cognition about what to do. Epistemic rationality pertains to epistemic cognition, and practical rationality pertains to practical cognition. Rationality pertains to “things the cognizer does” — acts, and in the case of epistemic rationality, cognitive acts. In particular, epistemic rationality pertains to “believings”. Epistemic justification pertains instead to beliefs — the products of acts of believing. But there seems to be a tight connection. As a first approximation we might say that a belief is justified iff it is rational for the cognizer to believe it. Similarly, practical cognition issues in decisions, and we can say that a decision is justified iff it is the product of rational practical cognition.

It is a commonplace of epistemology that epistemic cognition is not simply practical cognition about what to believe. If anyone still needs convincing of this, note that the logical properties of epistemic cognition and practical cognition are different. For instance, if Jones, whom you regard as a reliable source, informs you that *P*, but Smith, whom regard as equally reliable, informs you that ~*P*, what should you believe? Without further evidence, it would be irrational to decide at random to adopt either belief. Rather, you should withhold belief. Now contrast this with practical cognition. Consider Buridan’s ass, who starved to death midway between two equally succulent bales of hay because he could not decide from which to eat. That was irrational. He should have chosen one at random. Practical rationality dictates that ties should be broken at random. By contrast, epistemic rationality dictates that ties should not be broken at all except by the input of new information that renders them no longer ties.

So epistemic cognition and practical cognition work differently. And of course, there are many other differences between them — this is just one simple example of the difference.

On the other hand, a common presumption in epistemology is that epistemic justification is a *sui generis* kind of justification entirely unrelated to practical cognition, and one can study epistemic rationality and epistemic justification without ever thinking about practical cognition. One of the burdens of this paper will be to argue that this is wrong. I will argue that for sophisticated cognitive agents like human beings, an account of epistemic rationality must presuppose an account of practical rationality. I will defend this by discussing how epistemic cognition and practical cognition are intertwined. I will suggest that epistemic irrationality always derives from a certain kind of practical irrationality, and I will give an account of why we are subject to that kind of practical irrationality. It turns out that for what are largely computational reasons, it is desirable to have a cognitive architecture that, as a side effect, makes this kind of irrationality possible. This will be an important ingredient in an account of epistemic rationality, and it will explain why it is possible to hold beliefs unjustifiably, or more generally to be epistemically irrational.

3. Rationality and Reflexive Cognition

First a disclaimer. One way people can behave irrationally is by being broken. If a person suffers a stroke, he may behave irrationally. But this is not the kind of irrationality I am talking about. Philosophers have generally supposed that people don't have to be broken to be irrational. So when I speak of irrationality in this paper, I am only concerned with those varieties of irrationality that arise in intact cognizers.

The key to understanding rationality is to note that not all aspects of cognition are subject to evaluation as rational or irrational. For instance, visual processing produces a visual representation of our immediate surroundings, but it is a purely automatic process. Although the visual representation can be inaccurate, it makes no sense to ask whether it was produced irrationally. We have this odd notion of having control over certain aspects of our cognition and not over other aspects of it. We have no control over the computation of the visual image. It is a black box. It is both not introspectible and cognitively impenetrable. But we feel like we do have some control over various aspects of our reasoning. For example, you are irrational if, in the face of counter-evidence, you accept the visual image as veridical. The latter is something over which you do have control. If you note that you are being irrational in accepting that conclusion, you can withdraw it. In this sense, we perform some cognitive operations "deliberately". We have voluntary control over them.[1]

To have voluntary control over something, we must be able to monitor it. So mental operations over which we have voluntary control must be introspectible. Furthermore, if we have voluntary control over something we must be able to decide for ourselves whether to do it. Such decisions are performed by weighing the consequences of doing it or not doing it, i.e., they are made as a result of practical cognition. So we must be able to engage in practical cognition regarding those mental operations that we perform deliberately. To say that we can engage in cognition about some of our mental operations is to say that we are *reflexive cognizers*. We have the

[1] Compare the discussion of freedom and spontaneity in cognition in McDowell (1994).

following central connection between rationality and reflexive cognition:

> Rationality only pertains to mental operations over which we have voluntary control. Such operations must be introspectible, and we must be able to engage in practical cognition about whether to perform them.

I will refer to such cognition as *voluntary cognition*. This need not be cognition that we perform deliberately, but we can deliberately alter its course. Rationality only pertains to voluntary cognition.

4. Q&I Modules

Next, another preliminary. In studying rational cognition, philosophers have often focused their attention on reasoning, to the exclusion of all else. But it is important to realize that much of our belief formation and decision making is based instead on various shortcut procedures. Shortcut procedures are an indispensable constituent of the cognitive architecture of any agent that must make decisions rapidly in unpredictable environments. I refer to these as *Q&I modules* (quick and inflexible modules). I have argued that they play a pervasive role in both epistemic and practical cognition (Pollock 1989, 1995). Consider catching (or avoiding) a flying object. You have to predict the trajectory. You do not do this by measuring the velocity and position of the object and then computing parabolic paths. That would take too long. Instead, humans and most higher animals have a built in cognitive module that enables them to rapidly predict trajectories on the basis of visual information. At a higher level, explicit inductive or probabilistic reasoning imposes a tremendous cognitive load on the cognizer. We avoid that by using various Q&I modules that summarize data as we accumulate it, without forcing us to recall all the data, and then makes generalizations on the basis of the summary (Pollock 1989, 119ff).

Although they make cognition faster, Q&I modules are often subject to various sources of inaccuracy that can be corrected by explicit reasoning if the agent has the time to perform such reasoning. Accordingly, our cognitive architecture is organized so that explicit reasoning takes precedence over the output of the Q&I modules when the reasoning is actually performed, and the agent can often learn to mistrust Q&I modules in particular contexts. For instance, we learn to discount the output of the Q&I module that predicts trajectories when (often by using that module) we can predict that the flying object will hit other objects in flight.

5. Practical Irrationality

I distinguished between practical rationality and epistemic rationality. By implication, we can distinguish between practical irrationality and epistemic irrationality. Practical irrationality is easier to understand. The explanation turns on the role Q&I modules play in practical cognition. Paramount among the Q&I modules operative in practical cognition is one that computes and stores evaluations of various features of our environment — what I call *feature likings*. Ideally, feature likings would be based on explicit computations of expected values. But they are often based on a form of conditioning instead (see my (1995), (2001), and (2006a) for more discussion of this). The advantage of such *evaluative conditioning* is that it is often able to produce evaluations in the absence of our having explicit beliefs about probabilities and utilities.

It estimates expected values more directly. But it is also subject to various sources of inaccuracy, such as short-sightedness. Thus, for example, a cognizer may become conditioned to like smoking even though he is aware that the long term consequences of smoking give it a negative expected value.

Decision-making is driven by either full-fledged decision-theoretic reasoning, or by some of the shortcut procedures that are an important part of rational cognition. If a decision is based on full-fledged decision-theoretic reasoning, then it is rational as long as the beliefs and evaluations on which it is based are held rationally. This is a matter of epistemic rationality, because what is at issue is beliefs about outcomes and probabilities. If the decision is based on a shortcut procedure, it is rational as long as it is rational to use that shortcut procedure in this case. And that is true iff the agent lacks the information necessary for overriding the shortcut procedure. So the cognizer is behaving irrationally iff he has the information but fails to override the shortcut procedure. For instance, a person might have a conditioned feature liking for smoking, but know full well that smoking is not good for him. If he fails to override the feature liking in his decision making, he is being irrational. This seems to be the only kind of uniquely practical irrationality (i.e., practical irrationality that does not arise from irrationally held beliefs). We might put this by saying that smoking is the stereotypical case of practical irrationality. The smoker is irrational because he knows that smoking has a negative expected value, but he does it anyway.

What makes it easy to understand practical irrationality is that all decision making has to be driven by something, and if the cognitive system is not broken, these are the only ways it can be driven.

Overriding shortcut procedures is something one can explicitly decide to do. One can engage in higher-order cognition about this, and act on the basis of it. So overriding shortcut procedures is, in the requisite sense, under the control of a reflexive agent. Is the agent who fails to override a shortcut procedure just not doing enough practical cognition? That does not seem quite right. The smoker can think about the undesirable consequences of smoking, and conclude that he should not smoke, but do it anyway. He did all the requisite cognition, but it did not move him. The problem is a failure of *beliefs* about expected values to move the agent sufficiently to overcome the force of the shortcut procedures. The desire to do something creates a strong disposition to do it, and the belief that one should not do it may not be as effective. Then one is making a choice, but one is not making the *rational* choice. Note, however, that one can tell that one is not making the rational choice. Some smokers may deny that they are being irrational, but they deny it by denying the claims about expected values, not by denying that they should do what has the higher expected value.

I have argued that uniquely practical irrationality always arises from a failure to override the output of Q&I modules, and I have illustrated this with a particular case — the failure to override conditioned feature likings. There is a large literature on practical irrationality,[2] and I do not have time to survey it here. My general focus will be on epistemic irrationality instead. I have not done a careful survey of cases of practical rationality, but I think it is plausible that uniquely practical irrationality always consists of the failure to override the output of Q&I modules. This source of practical irrationality seems to be a design flaw in human beings, probably deriving from the fact that Q&I modules are phylogenetically older than mechanisms for reasoning explicitly about expected values. It is important to retain Q&I modules in an agent

[2] See particularly Kahneman and Tversky (1982).

archiecture, because explicit reasoning is too slow. In particular, it is important to retain evaluative conditioning, because explicit reasoning about expected values is both too slow and requires too much experience of the world for us to base all our decisions on it. But it appears that evolution has done an imperfect job of merging the two mechanisms.

The upshot is that practical irrationality is easy to understand. My suggestion is that it all derives from this single imperfection in our cognitive architecture. When I turn to epistemic irrationality, I will argue for the somewhat surprising conclusion that it too derives from this same source.

6. Reflexive Epistemic Cognition

Epistemic irrationality consists of holding beliefs irrationally. We have seen that rationality only pertains to mental operations over which we have voluntary control. Such operations must be introspectible, and we must be able to engage in practical cognition about whether to perform them. But why would a cognitive agent be built so that it has voluntary control over some of its mental operations? Why not build it so that it follows the desired rules for cognition automatically? What I will now argue is that there are good reasons for building a cognitive agent so that it is capable of such reflexive epistemic cognition.

6.1 Re-ordering the Cognitive-Task-Queue

The simplest form of reflexive cognition is about how to order our cognitive tasks. We always have more cognitive tasks to perform than we can perform immediately. We have multiple new pieces of information to explore, multiple goals to reason about, and so on. We cannot pursue all of these at once, so they go on a queue — the *cognitive-task-queue*. They are prioritized somehow and taken off the queue and explored in order of their priority. There must be a default prioritization in order to make this work. However, we can decide to take things in a different order. For example, given two problems, one of which aims at achieving a more important result, the default prioritization might have us look at the more important one first. However, we may know from experience that we are very unlikely to solve the more important problem, and much more likely to solve the lesser problem. We may then decide to look at the lesser problem first. This seems to be a straightforward matter of comparing expected values. So this is a case in which practical cognition can intervene and alter the course of other aspects of cognition. What we are doing is engaging in practical cognition that results in re-ordering the cognitive-task-queue. We are deciding what to think about, and in what order to address problems. The ability to do this seems very important. It makes us more efficient problem solvers. For instance, when an experienced mathematician or scientist addresses a problem, he can draw on his experience for how best to proceed — what reasoning to try first. An equally bright college freshman may be unsuccessful at solving the problem primarily because he has no special knowledge about how to proceed and so must take things in the default order imposed on his cognitive-task-queue.

6.2 Refraining from Accepting a Conclusion

Consider a second kind of reflexive cognition. We can have an argument for a conclusion, and see nothing wrong with it, but *without giving up the premises*, we may refrain from accepting the conclusion either because we have independent reason for thinking it is wrong, or because we are simply suspicious of it. What is to be emphasized is that we can do this without giving up the premises, even if the argument purports to be deductive. The liar paradox is an example of this. In the case of the liar paradox, we have a purportedly deductive argument that looks correct — we do not see anything wrong with it — but we *know* the conclusion is incorrect. A very important feature of our cognitive architecture is that we are able to "back out of the problem" without solving it. We don't just go crazy, or believe everything. We "bracket" the reasoning, setting it aside and perhaps coming back to it later to try to figure out what is wrong. But notice that we don't have to come back to it. We can decide that it is not worth our time to try to figure out what went wrong.

The same thing holds when we are just suspicious of the conclusion of an argument. This happens in philosophy all the time. We come upon arguments whose conclusions we find hard to believe. We don't immediately accept the conclusions or reject the premises. We set the arguments aside. Often, when we return to them, we are able to pinpoint flaws.

This also happens in mathematics. In the course of trying to prove a theorem, I might "prove" something that just doesn't seem right. When that happens, I don't forge on. I set it aside and, if I care enough, try to figure out what went wrong. Something similar happens throughout science. Scientists get evidence for conclusions they find unbelievable, even though they cannot immediately see what is wrong with their data. They do not automatically accept the conclusions. They have the power to set them aside and examine them more carefully later.

If reasoning worked like the steering mechanism on a car, we would have to accept the conclusions of these arguments, but in fact we don't. It is clearly desirable for an agent to be so constructed that it can treat arguments in this more flexible manner. This indicates that the simple rule, "When you have a good argument for a conclusion, you should accept the conclusion", is not a genuine rule of rationality. It is perhaps best viewed as a default rule to be followed in the absence of reflexive cognition.

6.3 Errors in Reasoning

A third variety of reflexive cognition concerns the fact that we sometimes make errors in reasoning. For instance, mathematicians almost always make mistakes initially when trying to prove complex theorems. Far be it from being the most certain form of knowledge, the results of mathematical reasoning are often among the least certain. This can seem puzzling. If reasoning is a mechanical process, it seems that it should only go wrong when something in the system breaks. For example, in its current implementation, OSCAR cannot make mistakes in reasoning. However, mathematicians are not broken just because they make mistakes. What is going on?

Part of the explanation lies in the fact that mathematicians cannot hold the entire proof in working memory. They have to rely upon longer-term memory to know what the earlier steps of the proof accomplished.[3] Typically, they write the proof down as they go along, but they tend to write it down only sketchily, so they also have to rely upon memory to interpret their notes. All of this makes it possible for them to be mistaken about what they have already accomplished, and so they may make new inferences from things they have not actually established. Memory gives us only a defeasible reason for believing what is recalled. Thus the reasoning underlying

[3] See the discussion of this in chapter three of Pollock (1987) and Pollock & Cruz (1999).

mathematical proof construction is actually defeasible and reflexive.[4] It is reflexive because mathematicians are reasoning about what their earlier reasoning accomplished (i.e., appealing to memory of their earlier reasoning and reasoning defeasibly from that).

Another part of the explanation of mathematical error lies in the fact that mathematicians often "sketch" an argument rather than writing it out in full detail. The sketch asserts that certain things are inferable on the basis of other things without actually working through the argument. This involves some kind of pattern matching or analogical reasoning. It is desirable to allow a cognizer to form beliefs on the basis of such sketches, because actually constructing the full argument is tedious and usually not necessary. In effect, the cognizer is reasoning probabilistically (using the statistical syllogism) in meta-reasoning about the possibility of constructing arguments. So this can also be regarded as a form of reflexive cognition. He is reasoning about what reasoning he could do. Note that this alters the default course of cognition, because that would dictate that the agent not adopt the belief until an argument has actually been produced. This illustrates that the rule, "Don't hold beliefs for which you do not have good arguments", is not a correct rule of rationality. It is again at best a default rule pertaining to non-reflexive cognition.

When we decide whether to accept a conclusion on the basis of a sketch of an argument, this typically involves some deliberation on our part. If it is really important that we get it right, we may be more reluctant to accept it without working out the sketch in more detail. For instance, if we were just curious about whether the theorem is true, a rough sketch may satisfy our curiosity. But if we intend to publish the result or use it to build airplanes or nuclear reactors, we will take more care. This seems to be a matter of practical reasoning about the expected value of accepting the sketch.

Now notice something that will be very important when I discuss philosophical methodology. If an argument sketch is wrong, in order to retract the conclusion on the basis of its being wrong, we have to be able to tell that it is wrong. We check various ways of filling out the sketch, and when they don't work we conclude inductively that the sketch is wrong, i.e., that there is no correct argument that fills it out. For this to work, we have to be able to tell that the inference sketched is not licensed by particular explicit arguments constructed using our built-in inference rules, where those arguments are intended to fill out the sketch. So we have to be able to introspect what we did, and we must be able to check its conformance with our built-in inference rules.

It is important to recognize that sketching an argument is not something that only mathematicians do. We can sketch arguments in other contexts as well. Someone might think to himself, "Should I believe in God? Well, something had to create the universe, so there must be a God." Someone else might think, "Well, if there were a God, there wouldn't be so much evil in the world, so there can't be a God." As arguments, these are grossly incomplete, and notoriously difficult to complete. But they certainly can lead to belief. So argument sketches do not have to concern mathematics. We can have argument sketches about anything.

There seems also to be an important connection between sketching arguments and certain aspects of planning. Planning is often said to be "hierarchical" (Sacerdotti 1977). We plan for how to do something using high-level actions. For instance, if I want to get to LA my initial plan might be no more elaborate than "Fly to LA". But to execute this plan, I must plan *how* to fly to LA. I must plan when to go, what flight to take, how to

[4] I first gave this account of mathematical reasoning in my (1987).

get a ticket, etc. Hierarchical planning is a good cognitive strategy, because we often have to make decisions quickly without having the information we need to fill in the details. We may not get that information until later — maybe not even until we begin executing the plan. For instance, consider a plan for driving across an unfamiliar city on an interstate highway. We may plan which highway to take, and which turns to make, but we do not plan ahead about which lanes to drive in. We decide the latter as we drive, taking account of the flow of traffic around us.

When we adopt a high-level plan, we are assuming that we can fill in the details later as we get more information. But this is something we can only know inductively. And we often have to begin execution of a plan before we finish working out the details. We might model hierarchical planning as the construction and subsequent expansion of a argument sketch that there is a particular sequence of actions we can perform that will achieve the goal. Alternatively, we might pursue the opposite reduction and think of the construction of argument sketches as hierarchical planning for building a complete argument.

6.4 Forms of Reflexive Epistemic Cognition

The upshot is that there are good reasons for making an agent a reflexive cognizer — enabling the agent to engage in cognition that alters the default course of cognition in various ways. We have seen at least three kinds of reflexive cognition that humans can perform:

- re-order the cognitive-task-queue
- believe something on the basis of a sketch of an argument rather than a full argument
- refrain from accepting the conclusion of an argument

Although it is desirable for an agent to be able to cognize in these ways, this also opens the door to epistemic irrationality. That is the topic of the next section.

7. Epistemic Irrationality

Rationality only pertains to mental operations that are introspectible and subject to voluntary control. We have seen three examples of how reflexive cognizers make use of voluntary control. Now let us look at how each of these kinds of reflexive cognition can lead to irrationality, and see whether we can characterize the source of the irrationality.

7.1 Re-ordering the cognitive-task-queue

Having the ability to decide what to think about can result in our not thinking about things we should think about. Several familiar forms of irrationality derive from this.

Not thinking about problems for a theory

Given a cherished theory, it may occur to one that certain considerations might generate a problem for the theory (a defeater). There is a temptation to not think about these considerations, so as to avoid discovering that the theory is wrong. But not doing so is, at least often, irrational. The ability to refrain from thinking about the defeaters derives from the more general ability to re-order the cognitive-task-queue, so the same reflexive ability that is useful for some purposes enables us to be irrational in others.

Just why is it irrational to avoid thinking about possible problems for our theory? It is practical cognition that allows us to do this. We desire not to be proven wrong, and given that desire, perfectly correct practical reasoning leads to the conclusion that one way to achieve that is to avoid thinking about the possible problem. If it is irrational not to think about the considerations that might generate a problem for the theory then the practical cognition must be irrational. If the reasoning from the desire not to be proven wrong is correct, then the only way the practical cognition can be irrational is for it to be irrational to have the desire. Can that be irrational? We can throw light on this by noticing that there is a spectrum of cases in which we fail to think about the possible problem, and in some of these cases we are being irrational, but not in others.

The simplest case is one in which you do not think about the possible problem for your theory just because you do not have time to do so. You cannot do everything. That is why we have a cognitive-task-queue. Science and philosophy have to compete for attention not only with other abstract intellectual tasks, but also with doing the laundry and taking out the garbage. Things to think about go on the cognitive-task-queue, and may never be pulled off if other more important things keep intervening. You should not give up the theory just because you *might* be wrong. If you had time to think it through you might well be able to dismiss the problem. Perhaps, without thinking it through, you should hold the belief with a bit less conviction (i.e., the probability of there being a defeater might diminish your degree of justification), but you should not simply give it up. So your belief is not automatically unjustified just because you do not think about the possible problem.

In the first case, the reason you fail to think about the problem is that you do not have time, not that you are trying to avoid being refuted. Perhaps that is where irrationality comes in. But consider a second case. Suppose you are a Nobel Prize winning physicist, famous for constructing the theory that may be challenged by the problem. The award of the Nobel Prize has given you great prestige, and made you able to do good things unrelated to the truth of the theory (e.g., collect food for starving refugees from war-torn countries). If it were to become known that the theory is wrong you would be the subject of ridicule and no longer able to accomplish these good things. Furthermore, you might know that if you find that you are wrong, you will not be able to conceal it. In this case, you are being perfectly rational in not wanting to be proven wrong, and correspondingly rational in not thinking about the possible problem. If you are being rational in not thinking about the possible problem for the theory, it is presumably rational for you to continue to believe the theory. Is your belief in the theory justified? I would think so. After all, the problem is only a *possible* problem. This case is like the first case in that if you did think about the problem, you might well be able to dismiss it. So the fact that you avoid thinking about the possible problem in order to avoid being refuted does not automatically render your belief unjustified.

Why should it ever be irrational for you avoid thinking about the possible problem in order to avoid being refuted? After all, you desire not to be refuted, and other things being equal, it is rational to try to satisfy your desires. But it is clear that there are cases in which it is irrational. The problem in these cases must be that your desire not to be refuted is irrational. It is natural to suppose that, ordinarily, your desire not to be proven wrong derives from a desire not to *be* wrong (by a kind of evaluative conditioning). Not thinking about the difficulty subserves the end of not being proven wrong, but not the end of not being wrong. So the desire to not be proven wrong is rationally problematic unless it has some other justification (as in the case of the Nobel Prize winner). It is based on a conditioned feature liking that ought to be overridden by your knowledge that it does not actually contribute to the end of not being wrong. In this respect, it is like having the desire to smoke in the face of knowing that smoking is bad for you. So the epistemic irrationality derives from a practical irrationality.

In a case in which it is irrational for you to avoid thinking about the possible problem, but you do so anyway, is your continued belief in the theory irrational? That does not follow, any more than it follows in the other cases. You are doing something wrong — you should check out the possible problem — but given that you are not going to, it would not be reasonable to give up the theory. After all, you have a lot of evidence for it, and you haven't found a *real* problem, only a possible problem. As before, perhaps you should believe the theory with a somewhat lower degree of conviction, but it would be irrational to give it up on the basis of the mere possibility of a problem.

Even though it would be irrational to give the belief up, I am not comfortable saying that the belief is justified. Epistemic justification is about whether we "should" hold a belief. But our normative judgments often have a more complex structure that cannot be expressed in this simple terminology. In cases of irrationality, if asked whether I should hold a belief in certain circumstances, the answer might be, "You shouldn't be in those circumstances, but given that you are, you should hold the belief." This does not imply "You should hold the belief" simpliciter. If you should hold the belief only because it is the best thing you can do given that, irrationally, you got yourself into those circumstances, do we want to say that it is a justified belief? The term "epistemic justification" is a term of art, so we can use it however we want, but this strikes me as a peculiar way to use it. We might say that the belief is *justified on the basis of the evidence that has been considered*, but not justified simpliciter because more evidence should have been considered.

One of the main lessons to be learned from this example is that epistemic rationality pertains to more aspects of cognition than epistemic justification. We can talk about a cognizer being rational or irrational in believing something, but also in how he carries out searches for additional evidence and defeaters, and rationality may pertain to other aspects of cognition as well. These considerations impinge on our judgments of justification. Epistemic justification has generally be taken to be the central notion of epistemology, but perhaps that is a mistake. If our interest is in understanding rational cognition, then the central notion should be rationality. In most cases, we can say what we want to say by talking about justified belief. But there may be no clear way to make sense of epistemic justification so that it has enough structure to be useful in talking about complex cases — particularly cases of "contrary-to-duty rational obligations", which are about what you should do given that you are in circumstances you should not be in.

It might be suggested that we can avoid this difficulty by understanding justification as relative to the "available evidence", but I doubt that is a well-defined notion. If the available evidence is just that currently contained in working memory (i.e., what the cognizer is explicitly thinking about), it is too restrictive. But if it includes other beliefs, they will be stored in long term memory and must be retrieved before they can be used in cognition. Beliefs stored in long term memory can only be retrieved with varying degrees of difficulty, and retrieval can range from virtually instantaneous to a process that takes hours or days or may fail altogether on any particular occasion. An agent cannot be culpable for not taking account of information that he has not yet been able to retrieve from long term memory. So there is no obvious way to define a notion of "available evidence" in such a way that a sensible notion of "justified belief" can appeal to it.

Epistemic justification seems to make unproblematic sense in cases that are not contrary-to-duty. Then justified beliefs are those held on the basis of rational epistemic cognition. So "epistemic justification" gives us a useful but rather crude tool for talking about the rationality of epistemic cognition. But in more complicated cases, where we want to know whether a cognizer should retain a belief in circumstances he is in only because he was irrational, talk of epistemic justification may not make clear sense.

Wishful thinking

A familiar form of irrationality is wishful thinking. How is wishful thinking possible? It does not seem that you can make yourself believe something just by wanting it to be true. That is not what happens in wishful thinking. I suggest that it is again a matter of irrationally re-ordering the cognitive-task-queue. Consider an example. My teen-age daughter has gone to a high school football game, and it occurs to me to wonder whether she took a jacket. Without really thinking about it I conclude, "Oh, she must have." What is happening here? It seems I am briefly rehearsing something like the following argument: (1) Most reasonable people would take a jacket; (2) she is a reasonable person; (3) so she took a jacket. Where I am going wrong is in not thinking hard enough about defeaters for this argument. She is a teenager and wants to look cool, and she thinks wearing a jacket is not cool.

What is important about this example is that I am drawing the conclusion on the basis of an argument — not just because I want it to be true. The term "wishful thinking" is a misnomer. You cannot make yourself believe something just by wishing it were true. On the other hand, my wanting it to be true may prevent me from looking for defeaters. Believing the conclusion makes me feel good (or allows me to avoid feeling bad), so this is a reason for not doing anything that might make me disbelieve it.

The practical reasoning involved in this example seems, on the surface, to be rationally correct. Why shouldn't I do it? After all, I want to feel good. However, the desire to believe that my daughter took a jacket derives from a desire for that to be true. I care about my daughter's well being, and not just about feeling good. Having the belief is not conducive to the truth of what is believed, so this is a rationally criticizable desire. Again, what is wrong here is the practical irrationality of the reasoning leading me to not think about defeaters.

This case differs from the previous case in that this time the belief is intuitively unjustified. In both cases, irrational practical cognition results in our not considering possible defeaters, but in the previous case the belief was initially justified and then a potential defeater occurred. In the present case we adopt the belief on the basis of a defeasible argument supporting it. Whenever we do this, we should immediately consider whether we already have any readily available defeaters. That is a requirement of rationality, and it is what the irrational practical cognition is preventing.

Hasty inference

We often engage in what might be called "hasty inference". For example, I may think about a hypothesis, think briefly about considerations that favor its being true, and then conclude that it is true and quit thinking about it. This is another example of not searching for defeaters.

One variety of hasty inference that has been noted in epistemology (Goldman 1979) is hasty generalization. In hasty generalization, we often think of the cognizer as generalizing on the basis of inadequate evidence rather than ignoring defeaters. But what is wrong with that is that the ignored evidence may contain counter-instances of the generalization, so this is again a matter of not taking adequate account of defeaters. It is made possible by my deciding what to think about. It differs from the previous case, however, in that instead of intentionally avoiding defeaters I am being insufficiently careful in searching for them. A plausible suggestion is that such carelessness arises from a desire to be finished with a task (in this case, searching for defeaters) regardless of whether you have completed it correctly. Presumably, the desire to be finished with a task derives, by evaluative conditioning, from the desire to have it accomplished, however it is not genuinely conducive to the goal of having the task accomplished. So this is analogous to the observation that the desire to not find defeaters is not conductive to the achievement of the desire to be right. In both cases evaluative conditioning leads to desires that we ought to override, and we are being irrational if we do not. Again, the resulting belief is unjustified.

If this is right, then many cases of hasty inference and hasty generalization derive from the phenomenon we have already noticed — re-ordering the cognitive-task-queue in order to achieve irrational desires. However, this is not the only kind of hasty inference, as we will see next.

7.2 Ignoring Evidence

Enabling the agent to refrain from accepting a conclusion makes it possible for the agent to refuse to adopt a belief for which he has good reasons, just because he finds the conclusion repugnant. Thus a person may refuse to believe that he has a potentially fatal disease, and thereby refrain from seeking treatment. He doesn't want it to be true that he has the disease, and that causes him to want not to believe it. This is another case of practical irrationality, susceptible to an analysis similar to that I proposed for wishful thinking. The desire to not believe one has the disease presumably derives from a desire that it not be true, but not believing it does not make it false. Having the desire not to believe it turns on evaluative conditioning and involves a failure to override conditioned desires with explicit knowledge of the values of outcomes. Because refraining from believing he has the disease prevents him from seeking treatment, the expected value of refraining from believing is strongly negative, just as in the case of smoking. Hence it is irrational to ignore the evidence and refrain from holding the belief.

7.3 Inadequate Arguments

If we can accept a conclusion on the basis of a sketch of an argument, we can do so on the basis of an inadequate sketch. Consider an example. Suppose you conclude that God does not exist because there is so much evil in the world. You have what seems to be a relevant consideration for your conclusion, but you don't have a complete argument, and it is not obvious how to turn it into an argument. Thinking about it superficially, you think it must be possible to turn this into a good argument, and so without thinking about it further you accept the conclusion. Perhaps you are actually justified in thinking there is an argument that can be constructed, but suppose you are not. Then you are being overly hasty.

In a case like this, it seems that you have a weak inductive or analogical reason for thinking there is an argument that can be given. If you think no more about the matter, it seems to me you are justified in your conclusion, albeit weakly. But if you think just a little more about how to spell out the argument and fail to find a way to do it, that failure should constitute a defeater. We can distinguish several possibilities:

- You fail to accept the defeater (that you cannot fill out the argument) on the basis of the inductive argument supporting it, or you fail to retract the conclusion in light of the defeater, because you want to believe the conclusion. This is irrational, as above, and your belief is unjustified.
- You don't follow up on the matter because you have more pressing things to think about. Then your belief remains weakly justified.
- You fail to accept the defeater on the basis of the argument supporting it because, through intellectual laziness, you just don't think about the matter. This is another case of carelessness, and can be treated as above. Again, your belief is unjustified.

8. A Single Source for Irrationality?

I have argued that the only cognitive acts that can be assessed for rationality are those potentially under the control of reflexive cognition. I called this *voluntary cognition*. I have also shown how reflexive cognition can make possible at least the most commonly cited kinds of epistemic irrationality. I argued that all of the varieties of epistemic irrationality I surveyed derive from practical irrationality that occurs in the course of reflexive cognition, and I think this is true in general. So on this account, epistemic irrationality derives from practical irrationality that occurs in the course of reflexive cognition. And I have suggested, somewhat tentatively, that all practical irrationality may also derive from a single source — the failure to override Q&I modules.. This often takes the form of failing to override conditioned feature likings in the face of explicit knowledge about expected values. If this is right, it follows that all irrationality, practical or epistemic, derives from this single source. From this I want to draw some general conclusions about epistemic rationality and epistemic justification.

9. Epistemic Rationality and Practical Rationality

An understanding of epistemic irrationality can, obviously, be employed to give an account of epistemic rationality. Epistemic cognition is rational just in case it is not irrational. I remarked in section two that philosophers have commonly supposed epistemic justification to be analyzable independently of practical rationality. However, at least in those cases where the notion makes clear sense, justified beliefs are those held on the basis of rational epistemic cognition, and the characterization of rationality for epistemic cognition will be dependent, in part, on a characterization of rationality in reflexive cognition. If, for example, we fail to find a defeater for a bit of reasoning because we have irrationally re-ordered the cognitive-task-queue so as to avoid finding it, then we may not be justified in believing the conclusion of the argument. However, the irrationality involved in re-ordering the cognitive-task-queue is practical irrationality. We employ practical reasoning in deciding whether to re-order the cognitive-task-queue, and the irrationality of that practical reasoning is what makes our belief unjustified. Thus an analysis of epistemic justification cannot be given independently of an account of practical rationality.

On the other hand, we should not jump to the conclusion that there is no difference between epistemic rationality and practical rationality. As I illustrated in section two, they often have quite different logical properties. My conclusion is just that you cannot give an analysis of epistemic rationality without talking about practical rationality. The converse seems equally clear. For example, decisions are based in part on beliefs about your situation. If you hold the beliefs unjustifiably, then your decision is also unjustified. So the analyses of epistemic and practical rationality must form a unified package. We need a single theory characterizing both and saying how they are connected.

It also follows from the present account that much work in epistemology falls short of producing rules for rational cognition. Accounts of how to reason inductively, or how to reason about times and causes, or how to form beliefs on the basis of perceptual input, are not themselves complete rules of rationality. At best, they describe the rules the cognizer should follow in the absence of reflexive cognition. They are "default rules" for how to cognize, but a complete account of rational cognition must describe not only these rules, but also explain how they fit into the more comprehensive architecture for rational cognition that characterizes both how we should reason in the absence of reflexive cognition and also how we should reason and perhaps violate some of these default rules in the course of reflexive cognition.

10. Normative and Descriptive Aspects of a Theory of Rational Cognition

The most important consequence of this account of irrationality concerns the nature of theories of rational cognition. Let us begin with the question of whether such theories are normative or descriptive. The standard philosophical view has it that theories of rationality are normative, not descriptive. They are about how we *should* cognize rather than about how we *do* cognize. Philosophy constructs normative theories of how to cognize, and psychology investigates what we actually do. The two enterprises are supposed to be largely orthogonal.

But I think this standard view is wrong. On my view, theories of rational cognition are, by and large, theories of how cognition actually works. It is striking, despite all the philosophical talk about irrationality, how rarely people behave irrationally. Almost all of our cognition is rational. Given the preceding account of irrationality, we can see why this is the case. On that account, irrationality arises from a single respect in which the human cognitive architecture works less well than we might desire — we find it difficult to override conditioned feature likings. Thus rationality has a much more central role in cognition than the traditional philosophical model supposes. If we confine our attention to those parts of cognition that are subject to rational evaluation, cognition *just is* rational cognition, with the exception that it occasionally goes wrong because of our failure to override Q&I modules. The result is that, for the most part, rules governing rational cognition are just rules describing how cognition actually works. Insofar as cognition does not work that way, it is because of this one glitch in the system. Given a description of the rules for rational cognition and a description of the glitch, we have a complete description of voluntary cognition. Describing voluntary cognition is a descriptive enterprise, potentially under the purview of psychology or cognitive science. And the theory of rational cognition makes up the bulk of that descriptive theory. So it seems to follow that constructing a theory of rational cognition is a descriptive enterprise.

But two questions remain. First, we certainly think of theories of rational cognition as normative — they tell us what we *should* do. How is this to be explained if they are just descriptive theories? Second, I described irrationality as deriving from a glitch in the system. What makes it a glitch rather than just a feature (à la Microsoft)? Calling it a glitch implies a value judgment. It assumes that rationality is desirable, and hence the glitch is a bad thing.

I suggest that theories of rationality are both descriptive and normative. In fact, their normativity derives from the descriptive theory. The two observations I just made that were intended to illustrate the normative character of rationality also reflect descriptive aspects of our cognitive architecture. We are so constructed that we fail to override Q&I modules in the face of explicit knowledge of expected values, and hence to behave irrationally. But we are also constructed in such a way that when we recognize that we have done so, we have a disposition to form the desire to "correct" our cognition, i.e., to retract the resulting beliefs and decisions and do the relevant reasoning over again. That is, we have a conative disposition to put a negative evaluation on our behavior and engage in reflexive practical cognition about how to correct the situation. That we have such a conative disposition is just another descriptive feature of our cognitive architecture. The negative evaluation works to combat the conditioned feature liking and reinforces the tendency to behave rationally. This is the sense in which rationality is normative. Of course, sometimes this mechanism is not strong enough to overcome the conditioned feature liking, as in the case of compulsive smokers. But all of this is just further description of how our cognitive architecture works. The normativity of rationality merely reflects our tendency to engage in certain kinds of reflexive cognition, and that is an entirely descriptive matter.

If theories of rational cognition are simply descriptions of certain aspects of our cognitive architecture, does that mean we should leave them to the psychologist to construct and confirm? No. First, these are matters that psychologists should, in principle, be able to investigate, but at this point psychologists do not have a good handle on how to do that. Second, it is a feature of our cognitive architecture, employing reflexive cognition as it does, that we have privileged access to the course of our rational cognition. Only voluntary cognition is rational, and what makes it voluntary is that we can monitor it introspectively, reason about it, and alter its course. So in order for cognition to count as rational, it must be such that we can in principle keep track of what we are doing. We also have to be able to recognize cases in which various kinds of failure occur. In particular, we must be able to discover that an argument sketch cannot be filled out. To do that, we must be able to tell that a particular way of filling it out does not constitute a good argument. This has to be an ability that is built into our cognitive architecture. If we can keep track of what we do, and we can recognize particular cases in which we have made mistakes, then we have the kind of data we need for formulating and confirming theories about when our cognition is not mistaken. We can notice, for example, that we often reason in accordance with *modus ponens*, and we can confirm inductively that we do not regard such reasoning as mistaken. Thus we can confirm that *modus ponens* is a correct rule of reasoning. Of course, this is a particularly simple case. It has, for example, proven much more difficult to discover rules describing the kinds of inductive and probabilistic reasoning that we regard as uncriticizable. But this much is clear. First, the reasoning we perform in this enterprise is straightforward scientific inference to the best explanation. Second, the data to be explained are what we can broadly call "introspective", and pertain both to what the actual course of our cognition is and to the judgments we are built to make about some of its correctness. There is nothing *a priori* about this enterprise.

11. Voluntary Cognition

Rationality only pertains to voluntary cognition. But in what sense is cognition voluntary? Voluntary actions are those driven by practical cognition. Understanding free will in the context of a deterministic system is a notoriously difficult problem. But it is not one we have to solve here. The point is that, however it is to be analyzed, there are things that we do voluntarily. That is what practical cognition accomplishes. We decide to do various things by engaging in practical cognition about what to do. So voluntariness pertains to practical cognition.

It is useful to think of cognition as a virtual machine implemented on our neurological substructure, much as your word processor is a virtual machine implemented on your computer's electronic substructure.[5] We can think of epistemic cognition as a subsidiary virtual machine that, given various inputs, can run by itself. When it does so it is following the default rules of epistemic cognition. In an important sense, epistemic cognition is not voluntary. Where voluntary control comes in is that the epistemic cognition machine has various parameters that we can choose to reset. So the epistemic cognition machine is embedded in the more general practical cognition machine, and although epistemic cognition can run by itself, practical cognition can tweak it by deciding to interfere in various ways with its default operation. In doing this, practical cognition is also interfering with itself, because the output of epistemic cognition is input to practical cognition — choices are based, in part, on our beliefs about the world.

Machines that are able to alter their own behavior might seem puzzling, but in fact they are not all that unusual. The role of epistemic cognition within practical cognition is similar to the operation of your word processor within your computer's operating system. These are two virtual machines, and the word processor can be thought of as embedded in the operating system. The word processor normally runs by itself, taking input from the keyboard (this is mediated by the operating system) and displaying output on the monitor (this is also mediated by the operating system). But the operating system can intervene in the operation of the word processor. For instance, when memory is low it may prevent the word processor from opening new files. In doing this it is also intervening in its own operation because it is the operating system, not the word processor itself, that creates the window that displays the new file.

So the sense in which epistemic cognition is under voluntary control and hence subject to rational evaluation is simply that epistemic cognition is embedded in and, in various ways controlled by, the more general practical cognition machine. Voluntary control is just what the practical cognition machine does.

12. Conclusions

My strategy has been to throw light on rational cognition by examining irrationality. I argued that practical irrationality derives from a general difficulty we have in overriding Q&I modules. Epistemic irrationality is possible because we are reflexive cognizers, and hence practical irrationality can affect our epistemic cognition. The upshot is that one cannot give a theory of epistemic rationality or epistemic justification without simultaneously giving a theory of practical rationality.

[5] See Pollock (2006) for an extended discussion of virtual machines.

A consequence of this account is that a theory of rationality is a descriptive theory, describing contingent features of a cognitive architecture, and it forms the core of a general theory of voluntary cognition. Most of the so-called “rules for rationality” that philosophers have proposed are really just rules describing default (non-reflexive) cognition. It can be perfectly rational for a reflexive cognizer to break these rules.

But rationality is also normative. The normativity of rationality is a reflection of a built-in feature of reflexive cognition — when we detect violations of rationality, we have a tendency to desire to correct them. This is just another part of the descriptive theory of rationality.

Although theories of rationality are descriptive, the structure of reflexive cognition gives philosophers, as human cognizers, privileged access to certain aspects of rational cognition. Philosophical theories of rationality are really scientific theories, based on inference to the best explanation, that take contingent introspective data as the evidence to be explained.

References

Goldman, Alvin

1979 What is justified belief? In *Justification and Knowledge,* ed. George Pappas. Dordrecht: D. Reidel.

Kahneman, Daniel, and Amos Tversky

1982 *Judgment under Uncertainty: Heuristics and Biases.* Cambridge University Press.

McDowell, John

1994 *Mind and World*, Cambridge, MA: Harvard University Press.

Pollock, John

1987 *Contemporary Theories of Knowledge*. Lanham, Maryland: Rowman and Littlefield.

1989 *How to Build a Person: a Prolegomenon*. Cambridge, MA: MIT Press.

1995 *Cognitive Carpentry*, Cambridge, MA: MIT Press.

1997 “Procedural epistemology”, Terry Bynum and Jim Moor (eds.), *The Digital Phoenix: How Computers are ChangingPhilosophy*, Blackwell’s, 17-36.

2001 “Evaluative cognition”, *Nous*, **35**, 325-364.

2006 “What Am I? Virtual machines and mental states”, *Philosophy and Phenomenological Research,* forthcoming.

2006a *Thinking about Acting: Logical Foundations for Rational Decision Making*, New York, Oxford.

Pollock, John, and Joseph Cruz

1999 *Contemporary Theories of Knowledge*, 2nd edition, Lanham, Maryland: Rowman and Littlefield.

Sacerdotti, E. D.

1977 *A Structure of Plans and Behavior*. Amsterdam: Elsevier-North Holland.